1 Cosmology: a brief refresher course

1.1 Fundamental assumptions

- Standard model of cosmology is based on two fundamental assumptions:
  - On sufficiently large scales, the Universe is **isotropic** – i.e. there is no preferred direction
  - Our position in the Universe is not special – the **Copernican principle**
- Together, these two assumptions imply that on large scales, the Universe is **homogeneous**
- We have good observational evidence for isotropy. For example, consider the cosmic microwave background. Once we subtract off the dipole due to our motion through space, we find that the remaining background is very close to flat, with inhomogeneities only at the 10^{-5} level.
- Direct observation evidence for homogeneity is harder to come by, but large-scale galaxy surveys find results consistent with a transition to homogeneity on scales $\sim 100\,\mathrm{Mpc}$; see e.g. Scrimgeour et al. (2012, MNRAS, 425, 116).
- In addition to these two key assumptions, we also typically assume that general relativity (GR) is the correct theory of gravity on large scales.

1.2 The Friedmann-Robertson-Walker metric

- A key lesson from relativity: space and time are not distinct things; instead, we live in a four-dimensional **space-time**.
- The geometry of space-time is described by a metric tensor $g_{\mu\nu}$. Locally, we can choose a coordinate system such that space-time appears flat (i.e. locally, space-time is described by the Minkowski metric of SR).
- On large scales, we cannot *a priori* assume that space-time is flat. However, the large-scale homogeneity and isotropy of the Universe allow us to write the metric in a relatively simple form:

$$ds^2 = c^2 dt^2 - a^2(t)\left(dr^2 + f_K^2(r)\left[d\phi^2 + \sin^2\theta d\theta^2\right]\right). \tag{1}$$

  Here, (r, ϕ, θ) are polar coordinates, $a(t)$ is a (time-dependent) scale factor and f_K is a function describing the curvature of the Universe:

$$f_K(r) = \begin{cases} K^{-1/2}\sin\left(K^{1/2}r\right) & K > 0 \\ r & K = 0 \\ |K|^{-1/2}\sinh\left(|K|^{1/2}\,r\right) & K < 0 \end{cases} \tag{2}$$

- Homogeneity and isotropy allow only three possibilities for the spatial curvature of the Universe on large scales. The case $K > 0$ corresponds to **positive curvature**, the case $K < 0$ to **negative curvature**, and the special case $K = 0$ to **no curvature**, i.e. a spatially flat Universe. Current observations suggest that the $K = 0$ case best describes the Universe we inhabit.

- The scaling factor a in the FRW metric cannot depend on our location in the Universe, but the assumptions of homogeneity and isotropy do not prevent it from being a function of time. The case $\dot{a} > 0$ corresponds to a Universe which is expanding; the case $\dot{a} < 0$ to one which is contracting.

1.3 Redshift

- If a is not constant, then photons propagating to us from distant sources will be **redshifted** (if the Universe is expanding) or **blueshifted** (if the Universe is contracting). The fact that the light that we observe from distant galaxies is redshifted tells us that we live in an expanding Universe.

- Consider light emitted from some source which is **comoving** with the expansion of the Universe at a time t_e, and which is observed by a comoving observer at $r = 0$ at time t_o. For light, we know from relativity that $ds = 0$, and if the direction of propagation is purely radial then the angular terms also vanish. We therefore have

$$c\,|dt| = a(t)dr. \tag{3}$$

  The coordinate distance between the source and the observer is simply:

$$r_{eo} = \int_{t_e}^{t_o} dr = \text{constant}. \tag{4}$$

  Alternatively, we can write this as:

$$r_{eo} = \int_{t_e}^{t_o} \frac{cdt}{a(t)} = \text{constant}. \tag{5}$$

- If r_{eo} is constant, then $\dot{r}_{eo}$ must vanish. Therefore:

$$\frac{\mathrm{d}r_{eo}}{\mathrm{d}t} = \frac{c}{a(t_o)}\frac{\mathrm{d}t_o}{\mathrm{d}t_e} - \frac{c}{a(t_e)} = 0, \tag{6}$$

  which means that

$$\frac{\mathrm{d}t_o}{\mathrm{d}t_e} = \frac{a_o}{a_e}. \tag{7}$$

  In an expanding Universe, a time interval dt_e at the source is lengthened by a factor a_o/a_e by the time it arrives at the observer.

- If we now take dt to be the time elapsed during the propagation of a single period of our light wave, i.e. $dt = \nu^{-1}$, then it is easy to show that

$$\frac{\nu_e}{\nu_o} = \frac{\lambda_o}{\lambda_e} = 1 + z = \frac{a_o}{a_e}. \tag{8}$$

1.4 The Friedmann equations

- So far, we have not used any results from GR. Redshift is a consequence of the FRW metric, and the FRW metric is a consequence of our assumptions of isotropy and homogeneity. GR only enters the picture when we want to determine *how* the scale factor $a(t)$ evolves with time.

- If we assume that the Universe is filled with a perfect fluid with energy density $\rho(t)$ and pressure $p(t)$, and is described by the FRW metric, then we can use GR to derive the following equations describing the behaviour of the scale factor $a(t)$:

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{Kc^2}{a^2} + \frac{\Lambda}{3}, \tag{9}$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3p}{c^2}\right) + \frac{\Lambda}{3}. \tag{10}$$

  Here, Λ is a dimensionless quantity known as the **cosmological constant**.

- We can combine these two equations to yield a third equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(a^3\rho c^2\right) + p\frac{\mathrm{d}}{\mathrm{d}t}\left(a^3\right) = 0. \tag{11}$$

- These three equations are known as the **Friedmann equations**.

1.5 Cosmological parameters

- We can divide the matter content of the Universe into two forms: relativistic and non-relativistic (aka "radiation" and "dust").

- For relativistic particles, the pressure is related to the energy density via

$$p = \frac{1}{3}\rho c^2, \tag{12}$$

  while for non-relativistic particles, p is much smaller than ρ and hence it is a good approximation to set $p = 0$.

- From equation 11, we see that for non-relativistic matter we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(a^3\rho c^2\right) = 0 \tag{13}$$

  and hence

$$\rho(t) = \rho_0 a^{-3}, \tag{14}$$

  where ρ_0 is the present-day energy density and we have chosen distance units such that $a_0 = 1$.

- For relativistic matter we instead have

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(a^3\rho c^2\right) + \frac{\rho c^2}{3}\frac{\mathrm{d}}{\mathrm{d}t}\left(a^3\right) = 0, \tag{15}$$

$$a^3\frac{\mathrm{d}\rho}{\mathrm{d}t} + 4\rho a^2\frac{\mathrm{d}a}{\mathrm{d}t} = 0, \tag{16}$$

  and hence

$$\rho(t) = \rho_{r,0}a^{-4}. \tag{17}$$

- The energy density of non-relativistic particles is dominated by their mass, and hence varies with redshift only because the spatial density of the particles varies. On the other hand, relativistic particles lose energy at a faster rate due to the effects of redshift.

- At this point, it is convenient to write the Friedmann equations in a simpler form with the help of a number of dimensionless parameters that we will now introduce.

- First of all, we have the **Hubble parameter**:

$$H(t) \equiv \frac{\dot{a}}{a}. \tag{18}$$

  The value of $H(t)$ at the present time is given by

$$H_0 = 100h\,\frac{\mathrm{km}}{\mathrm{s\,Mpc}} = 3.22\times10^{-18}h\,\mathrm{s}^{-1}, \tag{19}$$

  with $h \simeq 0.70$, and is known as the **Hubble constant**.

- Next, we have the **critical density**

$$\rho_{\mathrm{crit}} \equiv \frac{3H^2}{8\pi G}. \tag{20}$$

  The present day value is written as $\rho_{\mathrm{crit},0}$ and is given by

$$\rho_{\mathrm{crit},0} = 1.88\times10^{-29}h^2\,\mathrm{g\,cm}^{-3}. \tag{21}$$

- By dividing the densities of non-relativistic and relativistic matter by the critical density, we obtain the dimensionless density parameters Ω_{m} and Ω_{r}:

$$\Omega_{\mathrm{m}} \equiv \frac{\rho_{\mathrm{m}}}{\rho_{\mathrm{crit}}}, \qquad \Omega_{\mathrm{r}} \equiv \frac{\rho_{\mathrm{r}}}{\rho_{\mathrm{crit}}}. \tag{22}$$

- We can rewrite the first of the Friedmann equations in terms of H, Ω_m and Ω_r as:

$$H^2 = H_0^2\left[\Omega_{\mathrm{r},0}a^{-4} + \Omega_{m,0}a^{-3} + \frac{\Lambda}{3H_0^2} - \frac{Kc^2}{a^2H_0^2}\right]. \tag{23}$$

- If we define two further dimensionless density parameters,

$$\Omega_\Lambda \equiv \frac{\Lambda}{3H_0^2}, \tag{24}$$

and

$$\Omega_K \equiv \frac{-Kc^2}{H_0^2} = 1 - \Omega_{\mathrm{r},0} - \Omega_{\mathrm{m},0} - \Omega_\Lambda, \tag{25}$$

then we can rewrite Equation 23 as

$$H^2 = H_0^2 \left[\Omega_{\mathrm{r},0} a^{-4} + \Omega_{m,0} a^{-3} + \Omega_\Lambda + \Omega_K a^{-2}\right]. \tag{26}$$

- Suppose that $\Omega_\Lambda = 0$. In that case, the value of Ω_K depends on the total matter density. If this is greater than the critical density, then $\Omega_K < 0$ and hence $K > 0$; in other words, the Universe is **closed**. On the other hand, if the total matter density is less than the critical density, then $\Omega_K > 0$, $K < 0$ and the Universe is **open** and has negative curvature. Finally, if the matter density is *exactly* equal to the critical density, then $\Omega_K = 0$, $K = 0$, and the Universe is flat.

- Including a non-zero cosmological constant complicates this simple picture slightly, but the basic principle is the same: the global geometry of the Universe depends on the sign of Ω_K and hence on the value of $1 - \Omega_{\mathrm{m}} - \Omega_{\mathrm{r},0} - \Omega_\Lambda$.

- An important point to note here is that the terms on the right-hand side of Equation 26 have different dependencies on a and hence evolve at different rates. At the present-day, $\Omega_{\mathrm{r},0}$ is dominated by the cosmic microwave background, and $\Omega_{\mathrm{r},0} \ll \Omega_{\mathrm{m},0}$. Moreover, it also appears that in our Universe, $\Omega_K = 0$. Therefore, at the present time, the right-hand side of the Friedmann equation is dominated by the contributions from non-relativistic matter and the cosmological constant; in practice, $\Omega_{\mathrm{m},0} \simeq 0.275$, $\Omega_\Lambda \simeq 0.725$, and hence the cosmological constant term dominates.

- As we move to higher redshift, however, the size of the cosmological constant term does not change, while the non-relativistic matter term evolves as a^{-3}. Therefore, the latter dominates for $a < (\Omega_{\mathrm{m},0}/\Omega_\Lambda)^{1/3}$, corresponding to redshifts $z > 0.4$.

- In addition, since the radiation and matter terms also evolve at different rates, there will come a time when both are equal. This occurs at a redshift

$$z_{\mathrm{eq}} = \frac{\Omega_{\mathrm{m},0}}{\Omega_{\mathrm{r},0}} - 1, \tag{27}$$

known as the redshift of matter-radiation equality. Evaluating this, we find that $z_{\mathrm{eq}} \sim 5900$.

- In this course, we are mostly concerned with the range of redshifts between $z \sim 1000$ and $z \sim 20$, the so-called Dark Ages. In this range of redshifts, the matter term dominates, and we can write Equation 26 in the form

$$H^2 \simeq H_0^2 \Omega_{\mathrm{m},0} a^{-3}. \tag{28}$$

However, it is important to remember that this is an approximation that breaks down at lower and at higher redshifts.

- In the regime governed by this approximation – known as the **Einstein-de Sitter limit** – there is a simple relationship between the redshift z and the time t:

$$t = \frac{2}{3}\frac{1}{H_0\Omega_{\mathrm{m},0}^{1/2}(1+z)^{3/2}}. \tag{29}$$

1.6 Thermal history

- How does the temperature change as the Universe expands? The answer to this differs, depending on whether we are considering radiation or non-relativistic matter.

- We know that for a black-body radiation field, the energy density u_{rad} scales with temperature as $u_{\mathrm{rad}} \propto T^4$; this is the just the **Stefan-Boltzmann law**. Since the energy density of a radiation field scales as $u_{\mathrm{rad}} \propto (1+z)^4$, this implies that $T \propto (1+z)$, provided that the radiation field retains its black-body shape.

- Suppose that at redshift z_1, the Universe is filled with black-body radiation with a temperature T_1. A volume V_1 then contains

$$\mathrm{d}N_1 = V_1\frac{8\pi\nu_1^2\mathrm{d}\nu_1/c^3}{\exp\left(\frac{h\nu_1}{kT_1}\right)-1} \tag{30}$$

photons in the frequency range $\nu_1 \rightarrow \nu_1 + \mathrm{d}\nu_1$

- Provided that photons are not created or destroyed, but merely redshifted, the same set of photons at redshift z_2 occupy the frequency range $\nu_2 \rightarrow \nu_2 + \mathrm{d}\nu_2$, where $\nu_2 = f\nu_1$ and $\mathrm{d}\nu_2 = f\mathrm{d}\nu_1$, where $f = (1+z_2)/(1+z_1)$.

- The volume V_2 at z_2 corresponding to our original volume is given by $V_2 = V_1 f^{-3}$. Therefore, the number of photons $\mathrm{d}N_2$ in the frequency range $\nu_2 \rightarrow \nu_2 + \mathrm{d}\nu_2$ in a volume V_2 is simply $\mathrm{d}N_1$, and is given by

$$\mathrm{d}N_2 = \mathrm{d}N_1 = \frac{V_1}{f^3}\frac{8\pi f^3\nu_1^2\mathrm{d}\nu_1/c^3}{\exp\left(\frac{hf\nu_1}{kfT_1}\right)-1} \tag{31}$$

However, we can rewrite this as

$$\mathrm{d}N_2 = V_2\frac{8\pi\nu_2^2\mathrm{d}\nu_2/c^3}{\exp\left(\frac{h\nu_2}{kT_2}\right)-1}, \tag{32}$$

where $T_2 = fT_1$. Since we can apply the same argument to any frequency interval in our original spectrum, we see that an initial black-body spectrum retains its black-body shape and simply changes its temperature as the Universe expands.

- For a non-relativistic ideal gas, we can derive the evolution of the temperature from the relationship $p = K\rho^{\gamma}$, where γ is the adiabatic index. The expansion of the Universe is an adiabatic process, and hence K does not change as the Universe expands. Since we can write the pressure of an ideal gas in the form

$$p = \frac{\rho k T}{\bar{m}}, \tag{33}$$

  where $\bar{m}$ is the mean particle mass, this means that the relationship

$$\frac{\rho k T}{\bar{m}} = K\rho^{\gamma} \tag{34}$$

  must continue to hold as the Universe expands.

- From this, we see that the temperature of the gas must scale with the density as $T \propto \rho^{\gamma-1}$, and since the density evolves with redshift as $\rho \propto (1+z)^3$, this means that $T \propto (1+z)^{3\gamma-3}$. For an atomic (or ionized) gas with $\gamma = 5/3$, we therefore arrive at the result:

$$T \propto (1+z)^2. \tag{35}$$

- The temperature of a non-relativistic gas therefore falls off more rapidly than the radiation temperature, in the absence of any energy transfer between gas and radiation (or vice versa).

- In practice, the gas and radiation temperatures are strongly coupled at high redshift by Compton scattering. When photons Compton scatter off electrons, they may either lose or gain energy, depending on the details of the collision. However, we know from simple thermodynamics that in the limit of a large number of scatterings, energy will flow from the gas to the radiation field if the gas temperature $T_{\rm gas}$ is greater than the radiation temperature $T_{\rm rad}$, and from the radiation field to the gas if $T_{\rm rad} > T_{\rm gas}$.

- The energy transfer rate per unit volume can be written as

$$\Lambda_{\rm Comp} = \frac{4\sigma_{\rm T} a_{\rm sb} T^4 k n_{\rm e}}{m_{\rm e} c} \left(T - T_{\rm gas}\right), \tag{36}$$

  where $\sigma_{\rm T}$ is the Thompson scattering cross-section and $a_{\rm sb}$ is the Stefan-Boltzmann constant.

- If we compare this with the cooling rate due to the adiabatic expansion of the Universe, we find that Compton scattering dominates when

$$\frac{\Lambda_{\rm Comp}}{3nkT_{\rm gas}H(z)} > 1. \tag{37}$$

  Evaluating this, we find that Compton scattering dominates at redshifts greater than a few hundred.

- At high redshift, therefore, both gas and radiation temperatures evolve as $T \propto (1+z)$.

- In this course, we will mostly be concerned with the evolution of the Universe between redshifts $z \sim 1000$ and $z \sim 10$. However, it is useful at this point to remind you that there is considerable physics occurring at higher redshifts. In particular, primordial nucleosynthesis has already occurred.

- Protons and neutrons first form at the point where the temperature of the gas and radiation corresponds to $kT \simeq 1\,\mathrm{GeV}$.[1] At this point, the ratio of neutrons to protons is maintained in equilibrium by the conversion reactions

$$p + e^- \longleftrightarrow n + \nu \tag{38}$$

$$p + \bar{\nu} \longleftrightarrow n + e^+, \tag{39}$$

  where p represents a proton, n a neutron, e^- an electron, e^+ a positron, and ν and $\bar{\nu}$ are a neutrino and an anti-neutrino, respectively.

- Once the temperature of the Universe drops to $kT \simeq 800\,\mathrm{keV}$, these reactions "freeze-out" – the timescale associated with them becomes longer than the expansion timescale of the Universe. At the time that this happens, the neutron-to-proton number density is given by

$$\frac{n_\mathrm{n}}{n_\mathrm{p}} = e^{-\Delta mc^2/kT} \simeq \frac{1}{6}, \tag{40}$$

  where $\Delta mc^2 = 1.4\,\mathrm{MeV}$ is the mass difference between neutrons and protons.

- Although it is energetically favourable for the neutrons and protons to fuse together to form heavier nuclei, they cannot immediately do so, as immediately after freeze-out there are still too many extremely high energy photons around, and these photo-disintegrate any heavy nuclei that form. As the Universe expands and cools, however, the number density of these photons falls off exponentially, and once the temperature is $kT \simeq 80\,\mathrm{keV}$, heavier nuclei start to form in abundance. This occurs roughly three minutes after $t = 0$.

- The ratio of neutrons to protons at this point is around 1/7. It is smaller than the value at freeze-out because some of the neutrons have undergone beta decay.

- Almost all of these neutrons wind up in ^{4}He, while the vast majority of the remaining protons remain free. Small fractions of deuterium (i.e. ^2D), ^{3}He and lithium are also formed, but elements heavier than lithium form only in truly negligible amounts, owing to the lack of any stable nuclei with weights $A = 5$ or $A = 8$; the process of nucleosynthesis becomes "stuck" at helium, and cannot progress further.

- The precise abundances of the various nuclei depend on the details of the cosmological model, and in particular on the photon-to-baryon ratio. For the currently-favoured ΛCDM model, we have ***fill in figures here***

- At the point where we first enter the cosmological "Dark Ages" – the epoch of re-combination – the chemical composition of the gas is therefore roughly 76% ionized hydrogen, 24% ionized helium, and tiny traces of D, ^{3}He and Li.

[1] Immediately prior to this, the Universe was filled with a so-called quark-gluon plasma.

2 Recombination

2.1 The Saha equation

- As the Universe expands and the temperature drops, there eventually comes a point at which the reaction

$$\mathrm{H}^+ + \mathrm{e}^- \rightarrow \mathrm{H} + \gamma \tag{41}$$

  becomes much faster than its inverse, leading to the Universe transitioning from a mostly ionized to a mostly neutral state. This process is known as **recombination**.

- The simplest possible assumption that we can make about the process of recombination is that it takes place in thermodynamic equilibrium.

- Consider a small volume V, filled with N_e electrons, N_p protons and N_H hydrogen atoms. As we have already discussed, the temperature T in this volume is set by the interaction of the particles with the radiation field, which acts in this case as a large heat bath.

- If the timescale for the system to reach chemical equilibrium is much shorter than the expansion timescale of the Universe, then we can consider the reactions as taking place at constant volume and constant temperature. Thermodynamics tells us that in this case, chemical equilibrium is reached once the **Helmholtz free energy**,

$$F = U - TS, \tag{42}$$

  is a minimum.

- To see why, consider the first law of thermodynamics:

$$\mathrm{d}U = \mathrm{d}Q - p\mathrm{d}V. \tag{43}$$

  Here, $\mathrm{d}Q$ is the amount of heat exchanged with the surrounding heat bath. If the system is at constant volume, then the pressure work term vanishes and $\mathrm{d}U = \mathrm{d}Q$. From the second law of thermodynamics, we know that $\mathrm{d}S \geq \mathrm{d}Q/T$, and hence

$$\mathrm{d}U - T\mathrm{d}S \leq 0. \tag{44}$$

  If the temperature is constant, we can rewrite this as

$$\mathrm{d}\left(U - TS\right) \leq 0, \tag{45}$$

  or in other words, $\mathrm{d}F \leq 0$. Therefore, in any spontaneous reaction, F must stay the same or decrease, and equilibrium will only be reached once it reaches a minimum value.

- We can write the Helmholtz free energy for our system of electrons, protons and hydrogen atoms as

$$F = -kT \ln Z_c \tag{46}$$

where Z_c is the canonical partition function, given by

$$Z_c = \frac{Z_{\rm e}^{N_{\rm e}} Z_{\rm p}^{N_{\rm p}} Z_{\rm H}^{N_{\rm H}}}{N_{\rm e}! N_{\rm p}! N_{\rm H}!}, \tag{47}$$

where $Z_{\rm e,p,H}$ are the canonical partition functions for electrons, protons and hydrogen atoms, respectively.

- Taking the natural logarithm of Z_c yields

$$\ln Z_c = N_{\rm e} \ln Z_{\rm e} + N_{\rm p} \ln Z_{\rm p} + N_{\rm H} \ln Z_{\rm H} - \ln N_{\rm e}! - \ln N_{\rm p}! - \ln N_{\rm H}!, \tag{48}$$

but for a reasonable choice of volume, our numbers $N_{\rm e}$, $N_{\rm p}$ and $N_{\rm H}$ will be large enough for us to be able to use Stirling's approximation:

$$\ln N! \simeq N \ln N - N. \tag{49}$$

- Applying this, we find that:

$$\ln Z_c = N_{\rm e} \ln Z_{\rm e} + N_{\rm p} \ln Z_{\rm p} + N_{\rm H} \ln Z_{\rm H} - N_{\rm e}(\ln N_{\rm e} - 1) - N_{\rm p}(\ln N_{\rm p} - 1) - N_{\rm H}(\ln N_{\rm H} - 1). \tag{50}$$

- We now need to minimize F as a function of $N_{\rm e}$. After a bit of algebra, we find that:

$$\frac{\partial F}{\partial N_{\rm e}} = \ln Z_e + \ln Z_{\rm p} - \ln Z_{\rm H} - 2 \ln N_{\rm e} + \ln (N_{\rm B} - N_{\rm e}) = 0, \tag{51}$$

where $N_{\rm B} = N_{\rm p} + N_{\rm H}$ is the baryon number, and we have used the fact that $N_{\rm p} = N_{\rm e}$ and $N_{\rm H} = N_{\rm B} - N_{\rm e}$.

- Rearranging the above, we find that in equilibrium, we have:

$$\frac{N_{\rm p} N_{\rm e}}{N_{\rm H}} = \frac{Z_{\rm e} Z_{\rm p}}{Z_{\rm H}}. \tag{52}$$

- For a non-relativistic particle of mass m and chemical potential μ, the canonical partition function is

$$Z = \frac{gV}{(2\pi\hbar)^3} \int_0^\infty 4\pi p^2 e^{-(\epsilon-\mu)/kT} {\rm d}p \tag{53}$$

$$= \frac{gV(2\pi mkT)^{3/2}}{(2\pi\hbar)^3} e^{-(mc^2-\mu)/kT}, \tag{54}$$

where we have used the fact that $\epsilon = mc^2 + p^2/2m$.

- In chemical equilibrium, the total chemical potential vanishes, i.e. $\mu_e + \mu_p = \mu_H$. Also, the ionization potential of hydrogen can be written as $\chi = (m_e + m_p - m_H)c^2$.
- Using these results, we can write Equation 52 as:

$$\frac{x^2}{1-x} = \frac{(2\pi m_e kT)^{3/2}}{(2\pi\hbar)^3 n_B} e^{-\chi/kT}, \tag{55}$$

  where $x = n_e/n_B$ is the fractional ionization and $n_B = N_B/V$ is the number density of baryons. This equation is known as the **Saha equation**.
- Using this Equation, we can try to predict when the Universe recombines. If we set $x = 0.5$, and assume (following our discussion in the previous section) that $T = T_0(1+z)$, where T_0 is the present-day CMB temperature, then we can show that $z \sim 1400$.
- In reality, the ionization fraction drops to $x = 0.5$ at a slightly lower redshift, $z \sim 1300$. Clearly, something is wrong with our equilibrium argument.

2.2 The three-level atom

- The equilibrium approach that we took in the last section is simple, but wrong; it predicts too large a redshift of recombination. But why is it wrong?
- The reason is that we assumed that the radiation field would act simply as a heat bath, maintaining the temperature at a constant value (at any given redshift), but otherwise not acting to prevent the system from reaching chemical equilibrium.
- In reality, this is an over-simplification and the radiation field actually plays a very important role in governing the rate at which the electrons and protons can recombine.
- To see this, let's consider a simple three-level model of the hydrogen atom: we have our $1s$ ground state, the $2s$ and $2p$ excited states, and the ionized continuum. [**Sketch this at this point**].
- Suppose that we have recombination occurring directly into the ground state. The photon produced will have an energy $h\nu > 13.6\,\mathrm{eV}$, and if the local density of neutral hydrogen is sufficiently high, it will simply find another hydrogen atom to ionize, meaning that there is no net change in the number of neutral atoms.
- This phenomenon should be familiar to anyone who has studied the ISM. It prompts us to distinguish between two different cases when talking about recombination: case A, where the photons produced by recombination direct to the ground-state can escape from the region of interest, and case B, where the photons are re-absorbed close to their source.
- In the cosmological case, it is fairly obvious that case B applies. However, things are a little more complicated that this. Consider what happens when the atom recombines

into the $2p$ state. It will try to radiative de-excite from there to the ground state, emitting a Lyman-α photon in the process. However, the optical depth of the IGM to these Lyman-α photons is very large. Most will therefore be re-absorbed by other hydrogen atoms, leading to a large population of atoms in the excited state.

- At the epoch of recombination, the number of background photons with energy sufficient to ionize hydrogen is small; after all, this is why the gas recombines at this point. The number of photons capable of ionizing hydrogen from the $2p$ excited state, on the other hand, is still large. Therefore, recombination into the $2p$ state is followed, in the majority of cases, by re-ionization by these softer photons, rather than by a successful transition to the ground state.

- In the case of the $2s$ state, a similar argument applies. In this case, the lifetime of the state is long, as it is metastable – radiative de-excitation to the ground state occurs only via a two-photon emission process, which is forbidden, and hence has a small transition rate. An atom that is newly recombined into the $2s$ state is therefore far more likely to be re-ionized than to successfully reach the ground state.

- There is therefore a significant bottle-neck in the recombination process. The rate at which the number of neutral hydrogen atoms increases is set not by the rate at which recombinations occur into the excited states of hydrogen, but rather by two other processes: the rate at which atoms in the $2s$ decay to the ground-state via two-photon emission, and the rate at which Lyman-α photons are lost from the system by redshifting out of the line.

- We will now look at this in a more mathematical fashion. We begin by writing down the net production rate of hydrogen atoms per unit volume:

$$\alpha_{\rm e} n_{\rm e}^2 - \beta_{\rm e} n_{\rm 2s} = R + \Lambda \left(n_{\rm 2s} - n_{\rm 1s} e^{-h\nu_\alpha / kT} \right) . \tag{56}$$

- The first term on the left-hand side of this expression is the case B recombination rate, while the second represents ionization from excited states of the atom. We can write the latter in terms of the population of the $2s$ state rather than having separate terms for each excited state because the abundance of low energy photons is large enough to maintain the excited states in thermal equilibrium with respect to each other (although not with respect to the ground state); i.e. $n_{\rm 2p}/n_{\rm 2s} = 3$.

- In equilibrium, the right-hand side of this expression would be zero, implying that

$$\frac{\alpha_{\rm e}}{\beta_{\rm e}} = \left(\frac{n_{\rm 2s}}{n_{\rm e}^2} \right)_{\rm eqb} . \tag{57}$$

  However, the ratio on the right hand side is simply the Saha equation for the $2s$ state. We therefore have:

$$\beta_{\rm e} = \alpha_{\rm e} \frac{(2\pi m_{\rm e} kT)^{3/2}}{(2\pi\hbar)^3} e^{-B_2/kT}, \tag{58}$$

where $B_2 = 3.4\,\mathrm{eV}$ is the energy difference between the $2s$ state and the continuum. Note that since nothing on the right-hand side depends on the level populations of the hydrogen atoms, this relationship must hold even when we are not in thermal equilibrium.

- On the right-hand side of Equation 56, R represents the rate per unit volume at which Lyman-α photons are lost from the resonance, and $\Lambda = 8.23\,\mathrm{s}^{-1}$ is the Einstein coefficient for two-photon decay from the $2s$ state to the ground state. In addition to decays from $2s$ to the ground state we must also account for radiative excitation from the ground state to the $2s$ state, which we do by means of the second term in parentheses.

- To make further progress, we need to determine R. To help us do this, we first make a brief digression into the statistics of the radiation field. Recall that if we have a black-body spectrum, the radiation energy density per unit volume in the frequency interval $\nu \rightarrow \nu + \mathrm{d}\nu$ can be written as

$$u_\nu \mathrm{d}\nu = \frac{8\pi h\nu^3 \mathrm{d}\nu}{c^3} \frac{1}{e^{h\nu/kT} - 1}. \tag{59}$$

- The number density of photons in the same frequency interval then follows us

$$N_\nu \mathrm{d}\nu = \frac{u_\nu \mathrm{d}\nu}{h\nu} = \frac{8\pi\nu^2 \mathrm{d}\nu}{c^3} \frac{1}{e^{h\nu/kT} - 1}. \tag{60}$$

  This can alternatively be written in the form

$$N_\nu \mathrm{d}\nu = \frac{8\pi}{h^3} \left(\frac{h\nu}{c}\right)^2 \mathrm{d}\left(\frac{h\nu}{c}\right) \frac{1}{e^{h\nu/kT} - 1}, \tag{61}$$

$$= 2 \times \frac{4\pi p^2 \mathrm{d}p}{h^3} \frac{1}{e^{h\nu/kT} - 1}, \tag{62}$$

  where $p = h\nu/c$ is the photon momentum.

- In this expression, $4\pi p^2 \mathrm{d}p$ corresponds to the differential volume of momentum space for photons in this frequency range. Since h^3 is the elementary volume of phase space, the term $4\pi p^2 \mathrm{d}p/h^3$ therefore corresponds to the number of different states available in the frequency interval $\mathrm{d}\nu$ for a given photon polarization. Since there are two possible photon polarizations, it then follows that $2 \times 4\pi p^2 \mathrm{d}p/h^3$ is the total number of states available for this frequency interval.

- The factor

$$\frac{1}{e^{h\nu/kT} - 1} \tag{63}$$

  that relates the total number of states available to the number that are actually occupied is known as the **photon occupation number**, which we denote as $\mathcal{N}$. In the limit where $h\nu \gg kT$, we can write the value for a black-body radiation field as

$$\mathcal{N} \simeq e^{-h\nu/kT}. \tag{64}$$

- Within the Lyman-α resonance, the shape of the spectrum is not a black-body, owing to the influence of the recombination photons. On the short wavelength side of the resonance, we have a black-body spectrum, while within the resonance we have a step in the spectrum that connects smoothly up with the continuum to the long-wavelength side. **SIMON: sketch Figure 6.7 from Peebles here**.

- For simplicity, we assume that the spectrum within the resonance is flat. We justify this approximation by noting that the scattering time of the photons is much smaller than the Hubble time, so that each photon scatters many times (and is redistributed in frequency-space each time) before it is lost from the resonance.

- If we denote the photon occupation number within the resonance as $\mathcal{N}_\alpha$, and make use of the fact that during recombination, $h\nu \gg kT$, then we can use Equation 62 to write R as

$$R = \frac{2 \times 4\pi p^2}{h^3} pH \left(\mathcal{N}_\alpha - e^{-h\nu_\alpha / kT} \right). \tag{65}$$

  The Hubble parameter enters here because we can write the term $|\mathrm{d}p/\mathrm{d}t|$ as:

$$\left| \frac{\mathrm{d}p}{\mathrm{d}t} \right| = \left| \frac{\mathrm{d}p}{\mathrm{d}a} \right| \dot{a} = \frac{p}{a} \dot{a} = pH. \tag{66}$$

- If we introduce a new variable K, defined as

$$K \equiv \frac{\lambda_\alpha^3}{8\pi} H^{-1}, \tag{67}$$

  then we can write our expression for R in a simpler form:

$$KR = \mathcal{N}_\alpha - e^{-h\nu_\alpha / kT}. \tag{68}$$

- To close this set of equations, we need another expression for $\mathcal{N}_\alpha$. To derive this, we assume that the $2p$ and $1s$ levels are in statistical equilibrium: in other words, at any particular instant of time, the number of transitions from $2p$ to $1s$ is almost exactly balanced by the number of transitions from from $1s$ to $2p$, with the ratio between $1s$ and $2p$ changing only on a slower timescale as the Universe expands.

- In this case, we have

$$\left(B_{21} I_\alpha + A_{21} \right) n_{2\mathrm{p}} = B_{12} I_\alpha n_{1\mathrm{s}}, \tag{69}$$

  where B_{21}, B_{12} and A_{21} are the Einstein coefficients for the Lyman-α transition, I_α is the specific intensity of the radiation field in the Lyman-α resonance, and $n_{2\mathrm{p}}$ and $n_{1\mathrm{s}}$ are the number densities of hydrogen atoms in the $2p$ and $1s$ levels respectively. In writing down this expression, we have assumed that the influence of collisional excitation and de-excitation is negligible in comparison to the effect of the radiation field, which is a good approximation during the recombination epoch.

- The specific intensity I_α is related to the occupation number $\mathcal{N}_\alpha$ by

$$I_\alpha = \frac{2h\nu_\alpha^3}{c^2}\mathcal{N}_\alpha. \tag{70}$$

(This follows trivially from our definition of the occupation number). We can therefore write Equation 69 as

$$B_{21}\left(1+\mathcal{N}_\alpha\right)n_{2\mathrm{p}} = B_{12}\mathcal{N}_\alpha n_{1\mathrm{s}}, \tag{71}$$

where we have made use of the identity

$$A_{21} \equiv \frac{2h\nu_{21}^3}{c^2}B_{21} \tag{72}$$

to eliminate A_{21}.

- We therefore have

$$\frac{n_{2\mathrm{p}}}{n_{1\mathrm{s}}} = \frac{B_{12}}{B_{21}}\frac{\mathcal{N}_\alpha}{1+\mathcal{N}_\alpha}, \tag{73}$$

$$= \frac{g_2}{g_1}\frac{\mathcal{N}_\alpha}{1+\mathcal{N}_\alpha}, \tag{74}$$

where g_2 and g_1 are the statistical weights of the $2p$ and $1s$ levels, respectively.

- We know that the level populations of the $2p$ and $2s$ levels are in thermal equilibrium with respect to each other, and hence

$$\frac{n_{2\mathrm{p}}}{n_{2\mathrm{s}}} = \frac{g_2}{g_1}. \tag{75}$$

We therefore can write the ratio of the $2s$ and $1s$ level populations as

$$\frac{n_{2\mathrm{s}}}{n_{1\mathrm{s}}} = \frac{\mathcal{N}_\alpha}{1+\mathcal{N}_\alpha}. \tag{76}$$

- In thermal equilibrium, we know that when $h\nu \gg kT$, $\mathcal{N} \ll 1$. Since the radiation field in the resonance is close to a black-body, the same holds for $\mathcal{N}_\alpha$, which means that our ratio reduces to

$$\frac{n_{2\mathrm{s}}}{n_{1\mathrm{s}}} \simeq \mathcal{N}_\alpha. \tag{77}$$

- This is the final equation that we need in order to solve for R. We proceed by substituting this expression into Equation 56, which yields

$$\alpha_{\mathrm{e}}n_{\mathrm{e}}^2 - \beta_{\mathrm{e}}n_{1\mathrm{s}}\mathcal{N}_\alpha = R + \Lambda\left(\mathcal{N}_\alpha - e^{-h\nu_\alpha/kT}\right)n_{1\mathrm{s}}, \tag{78}$$

$$= \left(\mathcal{N}_\alpha - e^{-h\nu_\alpha/kT}\right) \times \left(\frac{1}{K} + \Lambda n_{1\mathrm{s}}\right). \tag{79}$$

- We can rearrange this to yield an expression for $\mathcal{N}_\alpha$:

$$\mathcal{N}_\alpha = \frac{\alpha_{\rm e} n_{\rm e}^2 + e^{-h\nu_\alpha/kT} \times \left(\frac{1}{K} + \Lambda n_{\rm 1s}\right)}{\frac{1}{K} + (\Lambda + \beta_{\rm e}) n_{\rm 1s}}, \tag{80}$$

$$= \frac{K\alpha_{\rm e} n_{\rm e}^2 + e^{-h\nu_\alpha/kT} \times (1 + K\Lambda n_{\rm 1s})}{1 + K(\Lambda + \beta_{\rm e}) n_{\rm 1s}}. \tag{81}$$

- Finally, we obtain the net recombination rate (the left-hand side of Equation 78) by substituting in our newly-derived value for $\mathcal{N}_\alpha$. After a little algebra, we find that

$$-\frac{{\rm d}n_{\rm e}}{{\rm d}t} = \left(\alpha_{\rm e} n_{\rm e}^2 - \beta_{\rm e} e^{-h\nu_\alpha/kT} n_{\rm 1s}\right) C, \tag{82}$$

  where

$$C = \frac{1 + K\Lambda n_{\rm 1s}}{1 + K(\Lambda + \beta_{\rm e}) n_{\rm 1s}}. \tag{83}$$

- The term $\beta_{\rm e} e^{-h\nu_\alpha/kT}$ is simply the photoionization rate from the ground state that one would derive using the Saha equation, and so the expression in parentheses is the one that we would use to describe recombination if we could ignore the effects of the Lyman-α resonance photons. The effects of these photons are accounted for in the **suppression factor** C.

- It is clear from its definition that $C < 1$; i.e. the resonance photons always delay recombination. Moreover, we see that if $\beta_{\rm e} \ll \Lambda$, then $C \simeq 1$, which makes physical sense: if the photoionization rate from the $2s$ level is much slower than the two-photon decay rate, then few excited atoms will be photoionized and the net recombination rate will be very close to the value that we would obtain if we just ignored the resonance photons.

- We can also look at whether two-photon decay or the loss of Lyman-α photons from the line via redshift is the more important effect. We can write the ratio between the two rates as

$$\frac{\text{Loss of Ly-}\alpha}{\text{Two-photon}} = \frac{R}{\Lambda \left(n_{\rm 2s} - n_{\rm 1s} e^{-h\nu_\alpha/kT}\right)}, \tag{84}$$

$$= \frac{1}{K} \frac{\mathcal{N}_\alpha - e^{-h\nu_\alpha/kT}}{\Lambda \left(\mathcal{N}_\alpha - e^{-h\nu_\alpha/kT}\right) n_{\rm 1s}}, \tag{85}$$

$$= \frac{1}{K\Lambda n_{\rm 1s}}. \tag{86}$$

- For our standard ΛCDM model, this is approximately $0.16/(1-x)$, where x is the fractional ionization. For $x \sim 1$, the loss of photons from the Lyman-α resonance is the dominant effect, but once recombination is well under way, two-photon decay from the $2s$ state dominates.

- Having looked in some detail at the beginning of recombination, it is natural to turn to the question of when recombination ends. The obvious answer to this question is that it ends once the gas has become fully neutral, but in practice this never happens.

- We can write the recombination timescale as

$$t_{\rm rec} = \frac{1}{k_{\rm rec,eff} n_0 (1+z)^3 x(z)}, \tag{87}$$

  where n_0 is the hydrogen number density at redshift zero and $x(z)$ is the fractional ionization of the gas at a redshift z, and $k_{\rm rec,eff}$ is the effective recombination coefficient (i.e. the value that one gets after accounting for the effects discussed above).

- As z and x decrease, this value rapidly increases. Eventually, it becomes longer than the expansion timescale of the Universe. Once we enter the regime where $t_{\rm rec} > t_{\rm H}$, which occurs at a redshift of around 700–800, we find little further evolution in the fractional ionization of the gas. The fractional ionization therefore **freezes out** at some non-zero value. The precise value depends on the baryon density parameter $\Omega_{\rm b}$ (which controls the size of n_0 in the expression for the recombination timescale), the total density parameter $\Omega_{\rm m}$ and the Hubble constant. The freeze-out value produced by our simple three-level model is approximately

$$x \simeq 1.2 \times 10^{-5} \frac{\Omega_{\rm m,0}^{1/2}}{h \Omega_{\rm b,0}}, \tag{88}$$

  which for our standard ΛCDM cosmological model yields $x \sim 2 \times 10^{-4}$.

- Finally, once we have computed $x(z)$ for all redshifts of interest, we can compute the optical depth of the Universe to Thomson scattering as a function of redshift:

$$\tau(z) = \int_0^l n(l) x(l) \sigma_{\rm T} {\rm d}l, \tag{89}$$

$$= \int_0^z n_0 \sigma_{\rm T} c \frac{x(z')(1+z')^2}{H(z')} {\rm d}z'. \tag{90}$$

- This expression gives the optical depth seen by a photon as it makes it way from redshift z to redshift 0. Any particular photon will have some redshift $z_{\rm ls}$ at which it scatters for the last time. We can use $\tau(z)$ to compute the probability distribution function for $z_{\rm ls}$, a quantity known as the **visibility function**:

$$p(z_{\rm ls}) = e^{-\tau(z_{\rm ls})} \left. \frac{{\rm d}\tau}{{\rm d}z} \right|_{z=z_{\rm ls}}. \tag{91}$$

- If we do this, then we find that $p(z)$ is well described by a Gaussian distribution, centred at $z \simeq 1100$ and with a standard deviation $\sigma_z \simeq 80$. This range of redshifts is known as the **last-scattering surface** (even though it is a volume, not a surface!), and the vast majority of the CMB photons that we see today scattered for the last time at this point. The CMB therefore tells us about the state of the Universe at the redshift of the last-scattering surface.

- Clearly, any uncertainty in $x(z)$ will affect $\tau(z)$ and hence will affect the position that we compute for the last-scattering surface. This in turn introduces uncertainty into the values of the cosmological parameters that we can derive using information from CMB observations. It is therefore important to understand $x(z)$ as accurately as possible.

2.3 Improved models of recombination

- The simplified three-level atom model discussed in the previous section is actually a pretty good approximation, and allows us to calculate the evolution of the fractional ionization with redshift to a precision of around 10%.

- For many purposes (e.g. understanding the later chemical evolution of the IGM), this is good enough. However, upcoming measurements of CMB fluctuations with e.g. Planck are sensitive to differences in the recombination history of the Universe at the 1% level.

- In order to improve on the three-level calculation, we need to add a lot more atomic physics. In this lecture, we will not discuss this additional physics in detail; instead, I will just briefly summarise some of the main effects, and some places to find out more.

- The most obvious improvement that we can make is to include many more levels of the hydrogen atom. Including additional levels changes our simple picture in two ways. First, atoms can recombine into a level with $n > 2$ and then radiatively cascade down to either the $2s$ or the $2p$ state, following which the routes to the ground state are the same as before. Second, atoms can recombine into an $n > 2$ level and then undergo a direct radiative transition to the ground state (e.g. via another of the Lyman series transitions).

- Recombination via the latter process is suppressed for the same reason that recombination via the Lyman-α transition is suppressed, but nevertheless some atoms will successfully recombine in this fashion owing to the loss of Lyman-series photons from the corresponding lines via the effects of redshift.

- Including a large number of additional levels turns the problem from one which can be attacked analytically, as above, to one which must be solved numerically. However, the computational requirements are not great: one must solve a coupled set of a few hundred ordinary differential equations, which can be done in a matter of minutes or less on a modern desktop computer.

- Another important correction comes from the inclusion of helium in the calculation. He^{++} recombines at a redshift $z \sim 6000$ and the Saha equation provides an adequate description of its recombination history. Because He^{++} recombines so early, it is essentially fully recombined by the time that hydrogen recombination gets started, and hence does not have any influence on the latter process.

- However, the same cannot be said for He^{+}. This recombines around $z \sim 2000$, and hence there is some overlap between the end of He^{+} recombination and the beginning of hydrogen recombination.

- The presence of He^+ influences the process of hydrogen recombination in two main ways. First, it increases the electron density present in the IGM, with the result that we can no longer assume that $n_e = n_{H^+}$. Second, photons produced during He^+ recombination can ionize neutral hydrogen. Specifically, the allowed transition from the 2^1P state to the 1^1S ground state produces a photon with an energy of 21.3 eV, well above the hydrogen photoionization threshold. In addition, the two-photon decay from the 2^1S singlet state to the ground state also often produces photons with energy greater than 13.6 eV.

- These two major pieces of additional physics were incorporated into calculations of hydrogen recombination by Seager et al. (2000, ApJS, 128, 407). They found that the main effect was to speed up recombination slightly. They also showed that this effect could be to a great extent be mimicked in a three-level calculation simply by increasing the effective recombination rate α_e by 14%. Finally, they made available a code called RECFAST that implemented all of this new physics. This is available at http://www.astro.ubc.ca/people/scott/recfast.html and has subsequently been updated a number of times to add further new physical effects.

- Calculations with RECFAST are sufficiently accurate that we can use the resulting recombination histories when analyzing WMAP data without worrying about introducing uncertainty; in this case, the observational uncertainties dominate. However, the original version of RECFAST was not accurate enough to allow us to do the same with Planck data, which will have much lower observational uncertainties.

- To achieve the sub-1% errors that are necessary if we are to make optimum use of the Planck data, a number of other effects must be accounted for. These include:
  - Stimulated two-photon decay from the $2s$ to $1s$ levels
  - Absorption of Lyman-α photons by the $1s$–$2s$ transition
  - Partial frequency redistribution within the Lyman-α resonance (i.e. the fact that the spectrum is not perfectly flat within the resonance)
  - Two-photon transitions from levels with $n > 2$
  - etc...

- Some of these processes have subsequently been incorporated into RECFAST. In addition, two other codes for computing recombination histories have become available: CosmoRec, written by Jens Chluba[2], and HyRec, written by Yacine Ali-Haïmoud and Chris Hirata[3].

- Further reading:
  - Peebles, 1968, ApJ, 153, 1 — *the original three-level calculation; a classic paper*
  - Seager et al., 2000, ApJS, 128, 407 — *the first $n \gg 3$ treatment*
  - Sunyaev & Chluba, 2009, AN, 330, 657 — *a good recent review*

[2]Available at http://www.cita.utoronto.ca/∼jchluba/Science_Jens/Recombination/CosmoRec.html
[3]Available at http://www.sns.ias.edu/∼yacine/hyrec/hyrec.html

3 Chemical evolution of pre-galactic gas

- Primordial gas contains only a few elements: most hydrogen and helium, small fraction of deuterium, tiny fraction of lithium. No significant quantities of elements heavier than lithium. Despite this, considerable chemical complexity is possible (e.g. Glover & Savin, 2009, MNRAS, 393, 911 construct a model of primordial chemistry tracking 30 species and 392 reactions).

- Much of this chemistry is only of academic interest. In this lecture, we focus on only a few key species and reactions.

- We start our survey with molecular hydrogen, H_2, the most common molecule in the Universe. The most obvious way to form this is via the radiative association reaction

$$H + H \rightarrow H_2 + \gamma. \tag{92}$$

  However, because H_2 has no dipole moment, this process is forbidden and proceeds at a negligibly slow rate.

- In the local Universe, large quantities of H_2 form via surface reactions on dust grains. However, primordial gas is dust-free, so this mechanism is not available there.

- Most of the H_2 that forms in the pre-galactic gas forms via one of two sets of ion-neutral reactions. The first of these involves the H^- ion as an intermediate:

$$\begin{aligned} H + e^- &\rightarrow H^- + \gamma, & (93)\\ H^- + H &\rightarrow H_2 + e^-. & (94) \end{aligned}$$

  The rate-limiting step in this **H^- pathway** is the formation of the H^- ion via a slow radiative association reaction. The subsequent reaction with atomic hydrogen occurs rapidly.

- The other main route to H_2 involves the H_2^+ ion:

$$\begin{aligned} H + H^+ &\rightarrow H_2^+ + \gamma, & (95)\\ H_2^+ + H &\rightarrow H_2 + H^+. & (96) \end{aligned}$$

  As before, the initial radiative association reaction is the rate limiting step.

- In both of these reaction chains, the H_2 formation rate depends on the fractional ionization of the gas. However, the reaction chains do not change this value: the electrons and protons essentially act as catalysts, and are not consumed.

- Once H_2 has formed in the pre-galactic gas, it is difficult to destroy. It has a binding energy of 4.48 eV, and hence direct collisional dissociation becomes ineffective for gas temperatures below a few thousand K, corresponding to redshifts $z < 1000$.

- Photodissociation of H_2 by the CMB is also relatively ineffective: direct photodissociation into the continuum is a very slow process, and in general the dominant process is two-step photodissociation via the electronically excited Lyman and Werner bands of H_2. **SIMON: draw figure here**.

- The lowest energy transition from the H_2 ground state to the Lyman state requires a photon energy of around 11.2 eV. The number density of CMB photons with this energy becomes completely negligible after the end of cosmological recombination, and at lower redshifts H_2 photodissociation is an unimportant process.

- Since destruction of H_2 in the pre-galactic gas is so difficult to bring about, the main thing limiting the amount that forms is the difficulty involved in making it. In particular, the amount of H_2 that forms at high redshifts is strongly limited by the effect of the CMB on the H^- and H_2^+ ions.

- The H^- ion has a binding energy of only 0.75 eV. The rate at which it is photodissociated by a black-body radiation field can be calculated, and is

$$R_{\mathrm{dis,H^-}} = 0.11 T_{\mathrm{r}}^{2.13} \exp\left(-\frac{8823}{T_{\mathrm{r}}}\right) \mathrm{s}^{-1}, \tag{97}$$

  where T_{r} is the radiation temperature. Comparing this with the rate of the associative detachment reaction (reaction 94), which is of order $10^{-9} n_{\mathrm{H}}\,\mathrm{cm}^3\,\mathrm{s}^{-1}$, we see that photodissociation dominates for redshifts greater than $z \sim 100$.

- The case of H_2^+ is somewhat similar, but is made more complicated by the fact that we need to account for the vibrational and rotational excitation of the H_2^+ ions. In the simple case in which all of the ions are in the ground state, the photodissociation rate for a black-body radiation field is

$$R_{\mathrm{dis,H_2^+,v=0}} = 2.0 \times 10^1 T_{\mathrm{r}}^{1.59} \exp\left(-\frac{82000}{T_{\mathrm{r}}}\right) \mathrm{s}^{-1}. \tag{98}$$

  Comparing this with the rate of reaction 96, $\sim 6 \times 10^{-10} n_{\mathrm{H}}\,\mathrm{cm}^3\,\mathrm{s}^{-1}$, we see that in this case, photodissociation dominates only at very high redshift, $z > 950$. This is much higher than the value we obtained for H^-, a result which is due to the higher binding energy of the H_2^+ molecular ion ($E_{\mathrm{D}} = 2.65\,\mathrm{eV}$).

- In practice, the assumption that all of the H_2^+ ions are in the ground state is not a good one. The ions formed by reaction 95 are typically formed in highly excited vibrational states, and states with $v \geq 1$ are also populated by radiative excitation of ground state H_2^+ molecules by the CMB. We can get an approximate idea of how this affects the survival of the H_2^+ ions by considering the simple case in which their rotational and vibrational levels have their local thermodynamic equilibrium (LTE) level populations. In this case

$$R_{\mathrm{dis,H_2^+,LTE}} = 1.63 \times 10^7 \exp\left(-\frac{32400}{T_{\mathrm{r}}}\right) \mathrm{s}^{-1}, \tag{99}$$

  and photodissociation dominates for $z > 330$. This is still significantly larger than the value for H^-, but much smaller than the $v = 0$ value for H_2^+.

- Detailed calculations that follow the populations of the individual H_2^+ levels show that even in this case, the photodissociation rate is underestimated, as the ions tend to have a super-thermal distribution, owing to the influence of reaction 95; see Coppola et al. (2011, ApJS, 193, 7) for more details.

- We therefore see that at high redshifts, rapid photodissociation limits the H^- and H_2^+ abundances to a very low level, and hence almost entirely suppresses H_2 formation, but that as we move to lower redshifts we will eventually reach a point at which photodissociation becomes unimportant. This occurs first for H_2^+. Once the photodissociation rate becomes small, the H_2^+ abundance climbs sharply, until it hits a new equilibrium value, set by the balance between H_2^+ formation by radiative association and H_2^+ destruction by charge transfer with atomic hydrogen:

$$H_2^+ + H \rightarrow H_2 + H^+. \tag{100}$$

- The peak value reached by the H_2^+ abundance is of order 10^{-12} and occurs shortly after photodissociation has become unimportant. At lower redshifts, the decreasing temperature of the gas decreases the H_2^+ formation rate but does not affect the temperature-independent destruction rate. **Simon: draw sketch of H_2^+, H_2 evolution, add to it as we go**.

- Most of the H_2 that forms via the H_2^+ pathway does so at this time. At lower redshifts, the decreasing gas density greatly lengthens the H_2 formation timescale, which quickly becomes longer than the Hubble time. The amount of H_2 formed by this mechanism is therefore quite limited. We can estimate the fractional abundance by finding the value for which $t_{\rm H_2,form} = t_{\rm H}$ at $z = 330$. The H_2 formation timescale for formation via the H_2^+ pathway is given by

$$t_{\rm H_2,form} = \frac{x_{\rm H_2}}{k_{\rm ct} x_{\rm H_2^+} n}, \tag{101}$$

  where $k_{\rm ct} = 6.4 \times 10^{-10}$ cm^3 s^{-1} is the rate coefficient for reaction 100. If we set $x_{\rm H_2^+} = 10^{-12}$ and $n = 10\,{\rm cm}^{-3}$ (approximately correct for $z \sim 330$), then we find that

$$t_{\rm H_2,form} \sim 10^{20} x_{\rm H_2}\ {\rm s}. \tag{102}$$

  In comparison, the Hubble time at this redshift is around $t_{\rm H} \sim 7 \times 10^{13}$ s. The amount of H_2 that forms before $t_{\rm H_2,form} = t_{\rm H}$ therefore corresponds to a fractional abundance of around 10^{-6}.

- Note that if we took $z \sim 1000$ as the point at which photodissociation became unimportant, as would be the case if all of the H_2^+ was in the ground state during the photodissociation-dominated epoch, then the value of the H_2 abundance that we would obtain from this argument would be at least a factor of $3^3 \sim 30$ larger. (In practice, the H_2^+ formation rate would also be larger, owing to the higher temperature, so the increase would be even larger than this, close to a factor of 100).

- Once we reach a redshift $z \sim 100$, a second phase of H_2 formation begins, driven by the H^- pathway. The peak H^- abundance is somewhat larger than the peak H_2^+ abundance, closer to 10^{-11}, owing to the faster formation rate of the H^- ion. However, the lower cosmological background density at $z \sim 100$ compared to $z \sim 330$ limits the amount of additional H_2 that can form. For the consensus ΛCDM model, the final H_2 abundance that we obtain if we treat the CMB as a perfect black-body is therefore around $x_{H_2} \sim 2 \times 10^{-6}$. As we shall see later, this is far less than we need to provide efficient cooling within the first protogalaxies.

- One additional complication that has not been greatly studied as yet is the role played by recombination photons. As we saw in our discussion of recombination, the CMB long-wards of the Lyman-α resonance is not a perfect black-body, as it is distorted by the Lyman-α photons emitted during the recombination epoch. Because the ratio photon-to-baryon ratio is very large (there are $\sim 10^9$ photons for every baryon), the distortion from a perfect black-body is small. Nevertheless, it proves to be important for some aspects of pre-galactic chemistry. In particular, it increases the photodissociation rate of H^- at $z < 100$ quite significantly compared to the rate that we would have for a perfect black-body, since there is no longer an exponential fall-off of the number density of photons with $E > 0.75\,\mathrm{eV}$.

- This effect was studied in detail by Hirata & Padmanabhan (2006, MNRAS, 372, 1175), who showed that it suppresses the amount of H_2 formed in the pre-galactic gas by a factor of a few.

4 Formation of structure: linear regime

- Up to this point in our discussion, we have assumed that the Universe is perfectly homogeneous on all scales. However, if this were truly the case, then we would not be here in this lecture theatre.

- We know that in reality, the Universe is highly inhomogeneous on small scales, with a considerable fraction of the matter content locked up in galaxies that have mean densities much higher than the mean cosmological matter density. We only recover homogeneity when we look at the distribution of these galaxies on very large scales.

- The extreme smoothness of the CMB tells us that the Universe must have been very close to homogeneous during the recombination epoch, and that all of the large-scale structure that we see must have formed between then and now.

- In this section and the next, we will review the theory of structure formation in an expanding Universe. We start by considering the evolution of small perturbations that can be treated using linear perturbation theory, before going on to look at which happens once these perturbations become large and linear theory breaks down.

4.1 Perturbation equations

- We start with the equations of continuity

$$\frac{\partial \rho}{\partial t} + \vec{\nabla} \cdot (\rho \vec{v}) = 0, \tag{103}$$

momentum conservation (i.e. Euler's equation)

$$\frac{\partial \vec{v}}{\partial t} + \left(\vec{v} \cdot \vec{\nabla}\right) \vec{v} = -\frac{\vec{\nabla} p}{\rho} + \vec{\nabla}\Phi \tag{104}$$

and Poisson's equation for the gravitational potential Φ:

$$\nabla^2 \Phi = 4\pi G \rho. \tag{105}$$

- We next split up the density and velocity in their homogeneous background values ρ_0 and $\vec{v}_0$ and small perturbations $\delta\rho$, $\delta\vec{v}$. If we let $\vec{r}$ represent physics coordinates, then our unperturbed velocity is simply

$$\vec{v}_0 = H\vec{r}, \tag{106}$$

i.e. it is the Hubble flow.

- To first order in our small perturbations, the continuity equation becomes

$$\frac{\partial \left(\rho_0 + \delta\rho\right)}{\partial t} + \vec{\nabla} \cdot (\rho_0 \vec{v}_0 + \delta\rho \vec{v}_0 + \rho_0 \delta\vec{v}) = 0. \tag{107}$$

This can be simplified by noting that the unperturbed density and velocity must also satisfy a continuity equation

$$\frac{\partial \rho_0}{\partial t} + \vec{\nabla} \cdot (\rho_0 \vec{v}_0) = 0. \tag{108}$$

Hence, our perturbation equation becomes

$$\frac{\partial \delta\rho}{\partial t} + \vec{v}_0 \cdot \vec{\nabla}\delta\rho + \rho_0 \vec{\nabla} \cdot \delta\vec{v} + \delta\rho \vec{\nabla} \cdot \vec{v}_0 = 0. \tag{109}$$

(Note that the $\nabla\rho_0$ term vanishes due to the homogeneity that we have assumed for our unperturbed state).

- If we define the density contrast

$$\delta \equiv \frac{\delta\rho}{\rho_0}, \tag{110}$$

then we can write this in a more compact form as

$$\dot{\delta} + \vec{v}_0 \cdot \vec{\nabla}\delta + \vec{\nabla} \cdot \delta\vec{v} = 0. \tag{111}$$

- From the momentum conservation equation, we obtain the relationship

$$\frac{\partial \delta\vec{v}}{\partial t} + \left(\delta\vec{v}\cdot\vec{\nabla}\right)\vec{v}_0 + \left(\vec{v}_0\cdot\vec{\nabla}\right)\delta\vec{v} = -\frac{\vec{\nabla}\delta p}{\rho_0} + \vec{\nabla}\delta\Phi, \tag{112}$$

  which can be simplified to

$$\frac{\partial \delta\vec{v}}{\partial t} + H\delta\vec{v} + \left(\vec{v}_0\cdot\vec{\nabla}\right)\delta\vec{v} = -\frac{\vec{\nabla}\delta p}{\rho_0} + \vec{\nabla}\delta\Phi, \tag{113}$$

- Finally, from the Poisson equation we have

$$\nabla^2\delta\Phi = 4\pi G\rho_0\delta. \tag{114}$$

- We now introduce comoving coordinates $\vec{x} \equiv \vec{r}/a$, and comoving peculiar velocities, $\vec{u} \equiv \delta\vec{v}/a$. Our spatial derivative transforms as

$$\vec{\nabla}_r = \frac{1}{a}\vec{\nabla}_x. \tag{115}$$

  Our time derivative, on the other hand, transforms as

$$\frac{\partial}{\partial t} + H\vec{x}\cdot\vec{\nabla}_x \to \frac{\partial}{\partial t}. \tag{116}$$

- In comoving coordinates, our perturbation equations become

$$\dot{\delta} + \vec{\nabla}\cdot\vec{u} = 0, \tag{117}$$

$$\dot{\vec{u}} + 2H\vec{u} = -\frac{\vec{\nabla}\delta p}{a^2\rho_0} + \frac{\vec{\nabla}\delta\Phi}{a^2}, \tag{118}$$

$$\nabla^2\delta\Phi = 4\pi G\rho_0 a^2\delta, \tag{119}$$

  where for simplicity we write $\vec{\nabla}_x$ simply as $\vec{\nabla}$.

- To close this set of equations, we also need an equation of state linking the pressure and density fluctuations:

$$\delta p = c_s^2\delta\rho = c_s^2\rho_0\delta. \tag{120}$$

4.2 Density perturbations

- By combining our first two perturbation equations, we can derive the following second-order differential equation of the density contrast:

$$\ddot{\delta} + 2H\dot{\delta} = \left(4\pi G\rho_0\delta + \frac{c_s^2\nabla^2\delta}{a^2}\right). \tag{121}$$

- To solve this, we start by decomposing δ into a set of plane waves:

$$\delta(\vec{x},t) = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\hat{\delta}\left(\vec{k},t\right)e^{-i\vec{k}\cdot\vec{x}}. \tag{122}$$

- Our Fourier amplitudes then obey the equation

$$\ddot{\hat{\delta}} + 2H\dot{\hat{\delta}} = \hat{\delta}\left(4\pi G\rho_0 - \frac{c_s^2k^2}{a^2}\right). \tag{123}$$

- In the limit $k \to 0$ (i.e. the long wavelength limit), this reduces to

$$\ddot{\hat{\delta}} + 2H\dot{\hat{\delta}} = 4\pi G\rho_0\hat{\delta}, \tag{124}$$

  which we recognise as the equation for a damped harmonic oscillator.

- In an $\Omega_m = 1$ Universe, we can write this equation as

$$\ddot{\hat{\delta}} + 2H\dot{\hat{\delta}} = \frac{3}{2}H^2\hat{\delta}. \tag{125}$$

  (In an $\Omega_m \neq 1$ Universe, things are a little more complex, but at our redshifts of interest, $\Omega_m \simeq 1$).

- We now try a solution of the form $\hat{\delta} \propto t^n$. This yields the equation

$$n(n-1)\frac{\hat{\delta}}{t^2} + 2Hn\frac{\hat{\delta}}{t} = \frac{3}{2}H^2\hat{\delta}. \tag{126}$$

  For $\Omega_m = 1$, we know that $H(t) = 2/3t$, and so this equation becomes

$$n(n-1)\frac{\hat{\delta}}{t^2} + \frac{4}{3}n\frac{\hat{\delta}}{t^2} = \frac{2}{3}\frac{\hat{\delta}}{t^2}. \tag{127}$$

  This equation has a non-trivial solution only when n satisfies

$$n^2 + \frac{n}{3} - \frac{2}{3} = 0. \tag{128}$$

  This equation has solutions $n = 2/3$ and $n = -1$, corresponding to a growing mode $\hat{\delta} \propto t^{2/3}$ and a decaying mode $\hat{\delta} \propto t^{-1}$.

- Convenient to express evolution of δ with redshift in terms of the current value, δ_0, and a term known as the linear growth factor, $D_+(z)$:

$$\delta(z) = \delta_0 D_+(z). \tag{129}$$

  For an Einstein-de Sitter Universe, $D_+(z) = (1+z)^{-1}$. For other cosmological models, we have the rather more complicated expression:

$$D_+(z) = \frac{1}{1+z}\frac{5\Omega_\mathrm{m}}{2}\int_0^1 \frac{\mathrm{d}a}{a^3H(a)^3}. \tag{130}$$

- Observations of the CMB show us that at $z \sim 1000$, the perturbations in the gas component have amplitudes that are of order 10^{-5}. If these perturbations then grow as $\hat{\delta} \propto t^{2/3}$, then by the time we reach redshift zero, they will have grown by at most a factor of 1000, and will still be of order 1%.

- Clearly, perturbations in the gas component alone cannot account for the highly inhomogeneous density distribution we see around us. So how does this structure form?

- **Dark matter** provides a resolution to this conundrum. Perturbations in the dark matter couple to the radiation field only through their gravitational influence (rather than by direct scattering, as is the case for the gas), and hence can be much larger than the gas perturbations without overly perturbing the CMB. By starting with much larger perturbations, we can reach the $\delta \sim 1$ regime much sooner, allowing us to form the observed structures.

4.3 Jeans length, Jeans mass

- From Equation 123, we see that the source term for our density perturbation equation is positive only if

$$k > k_{\rm J} \equiv \frac{2\sqrt{\pi G \rho_0}}{c_{\rm s}}. \tag{131}$$

  In other words, we will get growing perturbations only if they have wavenumbers that satisfy this criterion.

- An alternative way to express this criterion is in terms of a critical wavelength, defined as

$$\lambda_{\rm J} \equiv \frac{2\pi}{k_{\rm J}} = c_{\rm s}\sqrt{\frac{\pi}{G\rho_0}}. \tag{132}$$

  This critical value is known as the **Jeans length**. Only perturbations with wavelengths greater than the Jeans length will grow.

- Physically, we can understand the existence of this critical length scale by considering the balance between gravity and thermal pressure. If we take a small part of the pre-galactic gas and perturb it adiabatically, its density and temperature will increase. It will therefore be over-pressured relative to the surrounding gas, and the pressure gradients that we have created will try to smooth out the perturbation. Our perturbation will survive and grow only if its **self-gravity** – i.e. the gravitational force acting on the perturbation due to the perturbation's own mass – is larger than the pressure forces acting to smooth out the perturbation.

- It should be plain that for very small perturbations, with very low masses, pressure will overcome gravity. Similarly, it should be clear that on very large scales, gravity will win. There must therefore be some intermediate scale at which we go from being pressure-dominated to being gravity-dominated. This critical scale is just the Jeans length.

- We can also define a critical mass scale to go along with our critical length scale. This mass scale is known as the **Jeans mass** and is given by[4]

$$M_{\rm J} = \frac{4\pi}{3}\rho_0 \left(\frac{\lambda_{\rm J}}{2}\right)^3 . \tag{133}$$

- What happens if instead of gas, we consider dark matter? Most viable dark matter candidates are effectively collisionless, and hence have no sound speed *per se*. Does this mean that we can simply set $c_{\rm s} = 0$, and hence conclude that perturbations on all scales are unstable?

- For **cold dark matter** (CDM), this is actually a pretty good approximation. However, on very small scales it breaks down due to a phenomenon known as **free streaming**. This refers to the fact that our collisionless dark matter particles have a non-zero velocity dispersion. If their velocities are larger than the escape velocity of our perturbation, then they will simply stream away from the overdensity before it can undergo gravitational collapse.

- A careful analysis of this phenomenon leads one to derive an expression very similar to that for the Jeans length, only with the velocity dispersion of the dark matter in place of the sound speed. However, for CDM, the velocity dispersion is very small, and hence the Jeans mass and Jeans length of the dark matter are also very small; for instance, Diemand et al. (2005, Nature, 433, 389) show that for WIMP dark matter, the lowest mass dark matter halos should have masses of the order of an Earth mass.

4.4 Perturbations in a radiation-dominated Universe

- Up to this point, we have implicitly been assuming that the Universe is matter dominated. However, our initial density perturbations come into existence during the inflationary epoch and hence spend the first part of their life growing during the radiation-dominated era.

- In principle, correct treatment of perturbation growth during the radiation-dominated era requires a relativistic treatment of the governing equations. In practice, provided we are dealing with small perturbations, a non-relativistic treatment suffices.

- If we ignore pressure gradients (i.e. consider scales much larger than the Jeans length), then the governing equation for the growth of density perturbations in the radiation-dominated case can be derived in a similar fashion to that in the matter dominated case

[4]Note that there is a certain arbitrariness in our choosing to compute the mass within a sphere of radius $\lambda_{\rm J}/2$, and not, say, a sphere of radius $\lambda_{\rm J}$ or a cube of side length $\lambda_{\rm J}$. Consequently, the Jeans mass is a somewhat fuzzy concept, and should best be thought of as simply giving us a guide to the critical mass of an unstable perturbation. In practice, for perturbations with $M \sim M_{\rm J}$, we generally need numerical simulations in order to determine the ability of the perturbation to collapse and the timescale on which this occurs, particularly if the latter is comparable to the current expansion timescale.

if we make the substitutions $\rho \to \rho + p/c^2$ in the continuity equation, and $\rho \to \rho + 3p/c^2$ in the Poisson equation. Using the fact that $p = \rho c^2/3$ for radiation, we find that

$$\ddot{\hat{\delta}} + 2H\dot{\hat{\delta}} = \frac{32\pi}{3} G\rho_0 \hat{\delta}. \tag{134}$$

- We can rewrite this equation in terms of the Hubble parameter as

$$\ddot{\hat{\delta}} + 2H\dot{\hat{\delta}} = 4H^2\hat{\delta}, \tag{135}$$

Since $H = 1/2t$ in the radiation-dominated era, we find that we again have two solutions: a growing mode with $\hat{\delta} \propto t$ and a decaying mode with $\hat{\delta} \propto t^{-1}$. (Note that in deriving these soluations, we have assumed that $\Omega = 1$. This is always a good approximation during the radiation-dominated era).

- In terms of the scale factor, our growth mode is $\hat{\delta} \propto a^2$; hence, long wavelength perturbations grow much faster with increasing a in the radiation-dominated era than in the matter-dominated era, where they evolve only as $\hat{\delta} \propto a$.

- This is for perturbations on scales large enough that pressure forces are irrelevant. What happens on smaller scales? In the case of perturbations in the radiation or in the baryons (which are strongly coupled to the radiation at this point), the behaviour is fairly clear. We define a Jeans length as before,

$$\lambda_{\rm J} = c_{\rm s}\sqrt{\frac{\pi}{G\rho_0}}, \tag{136}$$

but in this case our sound-speed $c_{\rm s} = c/\sqrt{3}$, where c is the speed of light, and so

$$\lambda_{\rm J} = c\sqrt{\frac{\pi}{3G\rho_0}}. \tag{137}$$

- If we compare this number to the Hubble radius, $r_{\rm H} = c/H$, we find that $\lambda_{\rm J}/r_{\rm H} = \sqrt{8\pi^2/9} \simeq 3$; in other words, perturbations on scales comparable to the size of the observable Universe are suppressed during the radiation-dominated era.

- What about the dark matter? This does not couple directly to the radiation, and hence does not feel the radiation pressure. However, the growth of perturbations on scales $\lambda \ll r_{\rm H}$ is nevertheless suppressed, for a reason that we will now explain.

- If we consider scales $r \ll \lambda_{\rm J}$, then we can ignore any perturbations in the radiation component and treat it simply as a flat background. In this limit, the equation describing the growth of perturbations in the dark matter then becomes

$$\ddot{\hat{\delta}} + 2H\dot{\hat{\delta}} = 4\pi G\rho_{\rm m}\hat{\delta}. \tag{138}$$

Since we are in the radiation-dominated regime, we can write $H^2 = 8\pi G(\rho_{\rm m} + \rho_{\rm r})/3$. If we now change variables to $y \equiv \rho_{\rm m}/\rho_{\rm r} = a/a_{\rm eq}$, we find (after considerable algebra) that

$$\hat{\delta}'' + \frac{2+3y}{2y(1+y)}\hat{\delta}' - \frac{3}{2y(1+y)}\hat{\delta} = 0, \tag{139}$$

where $\hat{\delta}' \equiv {\rm d}\hat{\delta}/{\rm d}a$.

- If we adopt the trial solution $\hat{\delta} = Cy + D$, then it is easy to demonstrate that this is a solution to the above equation, provided that $D = 2C/3$. Therefore, we can write the growing mode solution as

$$\hat{\delta} = C\left(y + \frac{2}{3}\right), \tag{140}$$

  which becomes independent of y in the limit $y \ll 1$.

- We therefore see that as long as we are in the radiation-dominated regime, our small-scale dark matter perturbations do not grow. Physically, we can understand this effect as follows: the growth rate of the small perturbations (driven by $\rho_{\rm m}$) is much slower than the expansion rate of the Universe (driven by $\rho_{\rm r}$), and so δ is frozen at an approximately constant value for as long as $\rho_{\rm r} \gg \rho_{\rm m}$. Note, however, that this only holds on scales smaller than $\lambda_{\rm J}$. On scales larger than the Jeans length for the radiation-dominated fluid, both $\rho_{\rm r}$ and $\rho_{\rm m}$ contribute to the growth rate of the perturbations, which therefore can still grow significantly during this epoch.

4.5 Power spectrum

- It is widely believed that the seeds of our density fluctuations were generated by quantum fluctuations occurring during the inflationary epoch. If so, then so long as it remains in the linear regime, the density contrast field δ has a very useful property: it is a homogeneous, isotropic Gaussian random field. Its statistical properties are therefore complete determined by only two numbers: its mean and its variance.

- Mass conservation implies that $\langle\delta\rangle = 0$, where the angle brackets denote a space average.

- The variance of δ is conveniently described in terms of the **power spectrum** $P(k)$:

$$\left\langle \hat{\delta}(\vec{k})\hat{\delta}^*(\vec{k}')\right\rangle \equiv (2\pi)^3 P(k)\delta_{\rm D}\left(\vec{k} - \vec{k}'\right), \tag{141}$$

  where $\delta_{\rm D}$ is the Dirac delta function.

- The initial perturbations, seeded by quantum fluctuations during the inflation epoch, are typically assumed to have a power spectrum

$$P_i(k) \propto k, \tag{142}$$

  known as the Harrison-Zel'dovich spectrum.

- However, this initial power spectrum is subsequently modified because perturbations on different scales k do not all grow by the same amount during the radiation-dominated epoch.

- As we saw in the previous section, modes with wavelengths $\lambda \gg r_{\rm H}$ grow as $\delta \propto a^2$ in the radiation-dominated era, and $\delta \propto a$ in the matter-dominated era. On the other hand, modes with $\lambda \ll r_{\rm H}$ do not grow during the radiation-dominated era, and then subsequently begin to grow as $\delta \propto a$ during the matter-dominated era.

- To allow us to examine the effects of this difference in growth rates, let us make two simplifications. We will assume that the behaviour of a given mode changes instantly once $\lambda = r_{\rm H}$, and we will also assume that the evolution of the Universe changes instantly from radiation-dominated to matter-dominated at the redshift of matter-radiation equality (i.e. the redshift at which $\rho_{\rm m} = \rho_{\rm r}$).

- In this simplified picture, modes which have $\lambda > r_{\rm H}$ throughout the radiation-dominated era evolve as $\delta \propto a^2$ throughout the radiation-dominated era, and then as $\delta \propto a$ in the matter-dominated era. On the other hand, modes for which $\lambda = r_{\rm H}$ at some point during the radiation-dominated era evolve initially as $\delta \propto a^2$, then "freeze" once $\lambda = r_{\rm H}$, and finally start to grow again as $\delta \propto a$ at redshifts $z < z_{\rm eq}$. **SIMON: SKETCH THIS**. Small-scale modes (with large k) therefore have their growth suppressed relative to large-scale modes (small k).

- To quantify this, we first consider the mode that has $\lambda = r_{\rm H}$ at $z = z_{\rm eq}$; we speak of this mode "entering the horizon" at this time. We can write the comoving wavenumber for this mode as

$$k_0 = a_{\rm eq}\frac{2\pi}{r_{\rm H}} = 2\pi\frac{H_0}{c}\sqrt{\frac{2\Omega_{\rm m,0}}{a_{\rm eq}}} = 2\pi\frac{H_0}{c}\Omega_{\rm m,0}\sqrt{\frac{2}{\Omega_{\rm r,0}}}. \tag{143}$$

- Next, consider some mode that enters the horizon at the point when the scale factor is $a_{\rm enter} < a_{\rm eq}$. Up to this point, this mode has grown at the same rate as the mode with wavenumber k_0, but during the period from $a_{\rm enter}$ to $a_{\rm eq}$, it does not grow. On the other hand, the larger mode continues to grow as $\delta \propto a^2$ during this period.

- At $a_{\rm eq}$, the smaller mode is therefore suppressed relative to the larger mode by a factor

$$f_{\rm sup} = \left(\frac{a_{\rm enter}}{a_{\rm eq}}\right)^2 = \left(\frac{k_0}{k}\right)^2. \tag{144}$$

- After we enter the matter-dominated regime, the relative size of the modes does not change (so long as we remain in the linear regime). Since the power spectrum scales as δ^2, the final power spectrum is therefore related to the initial power spectrum by:

$$P_f(k) \propto \begin{cases} f_{\rm sup}^2 P_i(k) & k > k_0 \\ \\ P_i(k) & k < k_0 \end{cases} \tag{145}$$

- If our initial power spectrum is the Harrison-Zel'dovich spectrum, we find that

$$P_f(k) \propto \begin{cases} k^{-3} & k > k_0 \\ \\ k & k < k_0 \end{cases} \tag{146}$$

  where we have made use of the fact that $f_{\rm sup} \propto k^{-2}$. **SIMON: SKETCH final P(k)**

- This behaviour of the power spectrum has important consequences when we come to consider the formation of highly non-linear structures.

4.6 Relative velocity of dark matter and baryons

- Prior to recombination, the baryons and the radiation are tightly coupled together by Compton scattering, which allows for efficient momentum transfer from one component to another.
- As already noted, an important consequence of this is that the effective sound-speed in this coupled fluid is very high: $c_{\rm s,eff} = c/\sqrt{3}$, where c is the speed of light.
- Another important consequence is the fact that small-scale perturbations in the baryonic component are smoothed away by an effect known as **Silk damping**.
- If we have an overdensity, then locally we will have a higher number density of photons than in the surrounding gas. These photons will try to diffuse away from the overdensity, in order to restore the photon number density to equilibrium. Because of the high optical depth of the Universe at this epoch, they will do this via radiation diffusion (i.e. each photon will execute a random walk away from its initial location). As they do so, they will drag the baryons along with them, owing to the strong momentum coupling between baryons and photons.
- We can write the photon mean free path as

$$\lambda_{\rm mfp} = \frac{1}{n_{\rm e}\sigma_{\rm T}}, \tag{147}$$

  where $\sigma_{\rm T}$ is the Thomson scattering cross-section. The diffusion coefficient is then given by

$$D = \frac{1}{3}\lambda_{\rm mfp} c, \tag{148}$$

  and the diffusion radius (i.e. the distance to which the photons diffuse in time t) is given by

$$r_{\rm D} \simeq \sqrt{Dt}. \tag{149}$$

- At recombination, $t \sim 10^{13}$ s and $n_{\rm e} \simeq 400\,{\rm cm}^{-3}$. Therefore, $\lambda_{\rm mfp} \simeq 1.2$ kpc and $r_{\rm D} \simeq 6.2$ kpc, where these distances are in *physical units*. In comoving units, the diffusion length corresponds to ~ 6 Mpc, and hence Silk damping will erase any perturbations in the baryon-photon fluid on scales smaller than this.
- On scales $r > r_{\rm D}$, perturbations survive. As we have seen, we can consider the linear perturbations on these larger scales to be built up of a superposition of sound waves. Detailed analysis of the behaviour of the perturbations in this regime shows that owing to the effects of constructive interference, we expect to get the largest effects on wavelengths that are harmonics of the horizon scale, i.e. $\lambda = \frac{1}{n}\frac{c}{H(z)}$, where n is an integer, provided that $\lambda > r_{\rm D}$.
- This is a strong prediction of the basic hot Big Bang model, and has been successfully confirmed – these so-called "acoustic oscillations" are responsible for oscillatory pattern that we see if we measure the strength of the CMB anisotropies on a range of different angular scales.

- Now, what happens once the Universe recombines? Clearly, $n_{\rm e}$ drops rapidly and hence the photon mean free path increases significantly. However, at the same time, the coupling between photons and baryons becomes much weaker, as the timescale on which the two components can exchange momentum becomes comparable to or greater than the expansion timescale. Therefore, at a redshift $z \sim 1000$, the photons and baryons **decouple**. Although some scattering events occur after this time, and there remains a transfer of energy from the photons to the baryons, the rate at which momentum is transferred becomes too small to significantly affect the mean momentum of the baryons, and perturbations in the photons and in the baryons no longer evolve in the same fashion.

- As a result, the sound speed of the baryons drops very sharply from $c/\sqrt{3}$ to $c_{\rm s} = \sqrt{kT/\mu m_{\rm H}}$, the usual thermal sound-speed of an ideal gas. The Jeans length in the baryons also drops sharply, and on small-scales the baryons start to fall into the small-scale potential wells created by the dark matter. The dark matter, of course, does not couple to the radiation, and hence the perturbations in this component are not affected by Silk damping. Therefore, the small-scale perturbations in the baryons are regenerated, thanks to the dark matter, while the radiation component remains smooth on these scales.

- All of the effects that I have described so far were understood by the late 70s and early 80s. However, in 2010, Tseliakovich & Hirata pointed out another consequence of the baryon-photon coupling that had previously been overlooked. Before decoupling, the baryon-photon fluid has a non-zero velocity relative to the dark matter, owing to the effect of the acoustic oscillations in the former. What Tseliakovich & Hirata realized was that the baryons would initially retain this relative velocity even after decoupling.

- Detailed calculations (e.g. Tseliakovich & Hirata, 2010, Phys. Rev. D, 82, 083520) show that at decoupling, the rms size of the relative velocity[5] is around $30\,{\rm km\,s^{-1}}$. This is very small compared to the sound-speed prior to decoupling, but is large compared to the sound-speed of the baryons after decoupling, which is ~ 5–$6\,{\rm km\,s^{-1}}$.

- The coherence length of this relative velocity is comparable to the Silk damping scale, i.e. a few comoving Mpc. On small scales, therefore, the motion of the gas relative to the dark matter can be modelled as a bulk velocity. The size of this velocity decreases as the Universe expands – as with any peculiar velocity, it falls off as $v_{\rm pec} \sim (1+z)$. However, the sound speed in the gas also falls off with decreasing redshift, initially as $c_{\rm s} \propto (1+z)^{1/2}$ in the regime where $T_{\rm gas} \simeq T_{\rm r}$, and then as $c_{\rm s} \propto (1+z)$ in the regime where $T_{\rm gas}$ evolves adiabatically.

- At $z \sim 100$ – approximately the redshift at which the behaviour of T changes – the rms streaming velocity is around $3\,{\rm km\,s^{-1}}$ and the sound-speed is around $1.7\,{\rm km\,s^{-1}}$, and so the streaming motions are still supersonic. They remain so at lower redshift, as from this point on both $c_{\rm s}$ and $v_{\rm pec}$ evolve similarly with redshift.

[5]Note that in a homogeneous, isotropic Universe, the *mean* streaming velocity must be zero, but the *root-mean-squared* (rms) streaming velocity need not be zero.

- The full effects of this bulk motion on the formation of structure remain to be explored, but one obvious effect will be to increase the effective Jeans mass of the gas by a factor

$$f_{\rm inc} = \left(\frac{v_{\rm pec}}{c_{\rm s}} \right)^3 \sim 10. \tag{150}$$

5 Formation of structure: non-linear regime

5.1 The spherical collapse model

- Our treatment above works well in the linear regime, when $|\delta| \ll 1$, but breaks down once $|\delta| \sim 1$, since at this point we are no longer dealing with small perturbations, and hence can no longer use the tools of linear perturbation theory.

- The evolution of the gas and dark matter in the so-called **non-linear** regime is very complicated, and in general we need to use numerical simulations, rather than analytical techniques, in order to follow it.

- However, there are a few useful approximate models that we can look at that give us some guidance as to the behaviour of the gas and dark matter in the non-linear regime.

- The particular example that we're going to look at here is known as the **spherical collapse** model.

- Consider a spherical overdensity with uniform internal density. As this perturbation is overdense, it will reach some maximum physical radius and then collapse due to its own self-gravity. We denote the metric scale-factor at which the perturbation reaches its turn-around radius as $a_{\rm ta}$, and the radius of the perturbation at this point as $R_{\rm ta}$. We then define dimensionless coordinates:

$$x \equiv \frac{a}{a_{\rm ta}}, \qquad\qquad y \equiv \frac{R}{R_{\rm ta}}. \tag{151}$$

- If we consider, for simplicity, an Einstein-de Sitter Universe, then we can write the Friedmann equation as

$$\frac{{\rm d}x}{{\rm d}\tau} = x^{-1/2}, \tag{152}$$

where $\tau \equiv H_{\rm ta} t$ and $H_{\rm ta} = H_0 a_{\rm ta}^{-3/2}$.

- The equation of motion for the radius of our sphere can be written as

$$\ddot{R} = -\frac{GM}{R^2}, \tag{153}$$

$$= -\frac{4\pi}{3} \rho_{\rm ta} R_{\rm ta}^3 \frac{G}{R^2}. \tag{154}$$

Converting from t to τ, and defining a new overdensity parameter ζ through the equation

$$\rho_{\rm ta} = \frac{3H_{\rm ta}^2}{8\pi G}\zeta \tag{155}$$

allows us to write this in a much simpler form:

$$\frac{{\rm d}^2 y}{{\rm d}\tau^2} = -\frac{\zeta}{2y^2}. \tag{156}$$

Note that ζ is simply the overdensity of our perturbation at turn-around with respect to the cosmological background at the same time, measured in units of $\rho_{\rm crit}$.

- To solve our equation of motion, we need to specify some boundary conditions. The obvious choices are

$$\left.\frac{{\rm d}y}{{\rm d}\tau}\right|_{x=1} = 0, \qquad\qquad y|_{x=0} = 0, \tag{157}$$

i.e. our perturbation starts with zero radius when $a = 0$ and reaches its maximum size when $a = a_{\rm ta}$.

- With these boundary conditions, and with the help of the Friedmann equation, we can obtain the following solution

$$\tau = \frac{1}{\sqrt{\zeta}}\left[\frac{1}{2}\arcsin(2y-1) - \sqrt{y-y^2} + \frac{\pi}{4}\right], \tag{158}$$

which cannot easily be inverted to give y in terms of τ.

- At turn-around, $x = y = 1$ and $\tau = 2/3$, which means that

$$\zeta = \left(\frac{3\pi}{4}\right)^2 \simeq 5.55. \tag{159}$$

- By symmetry, the time taken from turn-around to collapse must be the same as that taken from the start to turn-around, i.e. in the absence of pressure forces or any non-sphericity, our perturbation will collapse to a point at $\tau = 4/3$, corresponding to $x = 2^{2/3}$.

- If our perturbation had not begun to evolve non-linearly, but had simply continued to evolve following the linear solution, its overdensity at this point would be merely

$$\delta_{\rm c} = 2^{2/3}\delta_{\rm ta} \simeq 1.69. \tag{160}$$

- In reality, our perturbation will never be perfectly spherical; non-spherical motions will develop as the perturbation collapses and will eventually halt the collapse.

- We assume that after the collapse halts, the collapsed object – often referred to as a **dark matter halo**, assuming we're considering a perturbation in the dark matter – relaxes into a state of **virial equilibrium**. In this case, the virial theorem tells us that the magnitude of the potential energy of the halo is equal to twice its kinetic energy:

$$|W_{\rm vir}| = 2T_{\rm vir} \tag{161}$$

Energy conservation implies that the kinetic energy of the virialized halo must be equal to the difference between the potential energy at turnaround, $W_{\rm ta}$, and the potential energy of the virialized halo:

$$|W_{\rm vir}| - |W_{\rm ta}| = T_{\rm vir}. \tag{162}$$

Therefore,

$$|W_{\rm ta}| = T_{\rm vir}, \tag{163}$$
$$|W_{\rm vir}| = 2|W_{\rm ta}|. \tag{164}$$

Since the potential energy of a spherical perturbation of radius R scales as $1/R$, this implies that

$$R_{\rm vir} = \frac{R_{\rm ta}}{2}. \tag{165}$$

- We can use this result to solve for the overdensity of the perturbation with respect to the background density at the time that the collapsing perturbation first virializes. Two factors contribute to this: the perturbation has collapsed (and hence increased its density), and the Universe has expanded (and hence decreased its density). The resulting density contrast is given by

$$\Delta = \left(\frac{2^{2/3}}{1/2}\right)^3 \zeta = 32\zeta = 18\pi^2 \simeq 178. \tag{166}$$

- Up to this point, we have been assuming an Einstein-de Sitter cosmological model. A similar analysis in the case where $\Omega_{\rm m} \neq 1$ is possible, but requires us to solve the resulting equations numerically. However, the end result is not too different from the Einstein-de Sitter case. For example, for $\Omega_{\rm m,0} = 0.3$ and $\Omega_\Lambda = 0.7$, we find that at $z = 0$, $\Delta \simeq 100$.

- In reality, non-linear structures forming in the dark matter are unlikely to be perfectly spherical. Indeed, **N-body** simulations that model the full non-linear evolution of the dark matter (albeit with some finite mass resolution) show that much is located in mildly overdense filaments and sheets, with larger overdensities located within these structures, particularly at the intersection of filaments.

- These highly overdense regions typically have an ellipsoidal morphology, and are commonly referred to as **dark matter halos**. Halos that have masses that exceed the local effective Jeans mass of the gas can capture gas from their surroundings. If this

gas then cools and undergoes further gravitational collapse, then the formation of stars will be the end result. In other words, these dark matter halos are the locations in which galaxies form. It is therefore important to understand their properties and their abundance within the Universe.

- In practice, even though these dark matter halos are far less symmetric than the idealized perturbation that we have considered in this section, the results of the spherical collapse model provide a reasonable first approximation when discussing their properties. This simple model also gives us a basis for determining the number density of halos of a given mass that we expect to find in the Universe, as we will see in the next section.

5.2 The Press-Schechter mass function

- Ideally, we would like to be able to determine the number density of halos of a given mass – the halo **mass function** – as a function of redshift without going to all the trouble and expense of running a large N-body simulation.
- Fortunately, we can! There is a simple analytical argument that allows us to derive a mass function that turns out to be a reasonable approximation to the true mass function. This argument was first formulated by Press & Schechter in 1974, and the resulting mass function has become known as the **Press-Schecher mass function**.
- We start by assigning a length scale $R(M)$ to each halo of mass M via

$$R(M) = \left(\frac{3M}{4\pi \rho_{\rm cr}(z) \Omega_{\rm m}(z)} \right)^{1/3}. \tag{167}$$

(In other words, R is the radius of a uniform sphere filled with matter at the mean density that has a total mass M).

- We next consider the density contrast smoothed on this scale R. This is defined as

$$\bar{\delta}_R(\vec{x}) \equiv \int \mathrm{d}^3 y \delta(\vec{x}) W_R(\vec{x} - \vec{y}), \tag{168}$$

where $W_R(\vec{x} - \vec{y})$ is a suitably chosen **window function**.

- If the density contrast δ is a Gaussian random field, then so is the smoothed field $\bar{\delta}_R$. For a Gaussian random field, the probability of finding any particular value $\bar{\delta}$ at a point in space $\vec{x}$ is given by

$$p(\bar{\delta}) = \frac{1}{\sqrt{2\pi\sigma_R^2}} \exp\left[-\frac{\bar{\delta}^2(\vec{x})}{2\sigma_R^2} \right], \tag{169}$$

where σ_R^2 is the smoothed density variance

$$\sigma_R^2 = 4\pi \int_0^\infty \frac{k^2 \mathrm{d}k}{(2\pi)^3} P(k) \hat{W}_R^2(k), \tag{170}$$

and $\hat{W}_R$ is the Fourier transform of our window function.

- The fraction of all points that have a density contrast greater than $\delta_{\rm c}$ (the linear density contrast for spherical collapse) is then given by

$$F = \int_{\delta_{\rm c}}^{\infty} p(\bar{\delta})\,{\rm d}\bar{\delta}, \tag{171}$$

$$= \frac{1}{2}{\rm erfc}\left(\frac{\delta_{\rm c}}{\sqrt{2}\sigma_R}\right), \tag{172}$$

  where erfc is the complementary error function.

- The great insight of Press & Schechter was that this number could also be identified as the total mass fraction in halos of masses greater than or equal to M.

- Another way of thinking about this: in the unsmoothed linear density contrast field, any points that have $\delta > \delta_{\rm c}$ correspond to gas that is now in a collapsed structure. By smoothing the density contrast field, we filter out those points that are in structures with scales less than $R(M)$ or masses less than M; hence, whatever is left must be in structures with mass $\geq M$.

- The mass fraction in halos with masses in the range $M, M + {\rm d}M$ is simply $\partial F/\partial M$. To compute this, we use the fact that we can write $\partial/\partial M$ as

$$\frac{\partial}{\partial M} = \frac{{\rm d}\sigma_R}{{\rm d}M}\frac{\partial}{\partial \sigma_R}, \tag{173}$$

  and also use the identity

$$\frac{\rm d}{{\rm d}x}{\rm erfc}(x) \equiv -\frac{2}{\sqrt{\pi}}e^{-x^2}. \tag{174}$$

  We find

$$\frac{\partial F}{\partial M} = \frac{1}{\sqrt{2\pi}}\frac{\delta_{\rm c}}{\sigma_R}\frac{{\rm d}\ln\sigma_R}{{\rm d}M}\exp\left(-\frac{\delta_{\rm c}^2}{2\sigma_R^2}\right). \tag{175}$$

- If we integrate this over all masses, we find we have a normalization problem:

$$\int_0^{\infty}\frac{\partial F(M)}{\partial M}{\rm d}M = \frac{1}{2}. \tag{176}$$

  Press & Schechter dealt with this by (somewhat arbitrarily) multiplying the mass function by a factor of two. The actual resolution to this problem was recognized 17 years later by Bond et al. (1991, ApJ, 379, 440), and requires us to derive the mass function in a somewhat different fashion, using the methods of excursion set theory. However, this is outside the scope of the present course.

- Given the correctly normalized version of $\partial F/\partial M$, we can then compute the comoving halo number density simply by multiplying by ρ_0/M:

$$N(M,z)\,{\rm d}M = \sqrt{\frac{2}{\pi}}\frac{\rho_0\delta_{\rm c}}{\sigma_R}\frac{{\rm d}\ln\sigma_R}{{\rm d}M}\exp\left(-\frac{\delta_{\rm c}^2}{2\sigma_R^2}\right)\frac{{\rm d}M}{M}. \tag{177}$$

- The redshift dependence of this expression enters because σ_R increases as the Universe expands and the density perturbations grow. It is therefore often convenient to write the above Equation in terms of $\sigma_{\rm R,0}$, the variance of the linear density field at $z = 0$, and the linear growth factor $D_+(z)$. In this case, we have

$$N(M,z)\,{\rm d}M = \sqrt{\frac{2}{\pi}}\frac{\rho_0\delta_{\rm c}}{D_+(z)\sigma_{R,0}}\frac{{\rm d}\ln\sigma_{R,0}}{{\rm d}M}\exp\left(-\frac{\delta_{\rm c}^2}{2D_+(z)^2\sigma_{R,0}^2}\right)\frac{{\rm d}M}{M}. \tag{178}$$

- To help us understand the behaviour of this mass function, let us start by considering the simple case in which our power spectrum $P(k)$ is a power-law function of k, i.e. $P(k) \propto k^n$. In this case, $\sigma_{\rm R,0}$ is given by

$$\sigma_{R,0}^2 = 4\pi\sigma_N^2\int_0^\infty \frac{k^{2+n}{\rm d}k}{(2\pi)^3}\hat{W}_R^2(k), \tag{179}$$

  where σ_N is some appropriately chosen normalization factor that fixes the normalization of the power spectrum. We often choose to express this normalization in terms of σ_8, the value of σ at $z = 0$ within a sphere of radius $R = 8h^{-1}$Mpc.

- If we assume, for simplicity, that our window function is a top-hat in k-space, so that

$$\hat{W}_R = \begin{cases} 0 & k > 2\pi/R \\ 1 & k < 2\pi/R \end{cases} \tag{180}$$

  then we find that

$$\sigma_{R,0}^2 \propto \int_0^{2\pi/R} k^{2+n}{\rm d}k \propto R^{-3+n}. \tag{181}$$

  Since $R \propto M^{1/3}$, we therefore find that $\sigma_{R,0} \propto M^{-(3+n)/6}$.

- If we consider small scales, so that we can set the exponential term in our mass function equal to one, then we find that

$$N(M,z)\,{\rm d}M \propto M^{(n-9)/6}{\rm d}M. \tag{182}$$

  We saw in a previous lecture that $P(k) \propto k^{-3}$, and hence on small scales $n = -3$. We therefore find that at the low-mass end, the mass function scales as

$$N(M,z)\,{\rm d}M \propto M^{-2}{\rm d}M. \tag{183}$$

- We therefore see that there are many more low-mass halos than high-mass halos. Moreover, the mass found in each logarithmic mass bin is constant, demonstrating that these low-mass halos do not only dominate the number counts but also represent a significant fraction of the total available mass. This will have important consequences later on, when we consider the effects of feedback from early protogalaxies.

- At the high mass end of the mass function, the exponential term generally dominates. The presence of this term means that although, in principle, there is a non-zero probability of finding a halo of arbitrarily large mass at any given redshift, in practice the probability soon becomes so small that the chance of finding one within the observable Universe is tiny; i.e. we may as well consider it to be zero, for all intents and purposes.

- It is often useful to quantify the rarity of a given halo in terms of the argument of this exponential. For instance, suppose that we are interested in a halo with a mass such that

$$\frac{\delta_{\rm c}}{\sigma_{R,0} D_+(z)} = 3. \tag{184}$$

  Rearranging this expression, we find that

$$\sigma_{R,0} = \frac{1}{3}\frac{\delta_{\rm c}}{D_+(z)}, \tag{185}$$

  and hence in order to form such a halo, we need a local upwards fluctuation in the density contrast field that corresponds to a three-sigma fluctuation. We know from numerical integration of the Gaussian distribution that such a fluctuation occurs with a probability of around 1%, and hence around 1% of the total mass in the Universe is to be found in regions where δ is this large or larger.

- When we talk about the "first" objects of a given mass scale to form, we therefore need to be careful what we mean. Do we mean the very first object to form within the observable Universe? In that case, we are talking about something that is approximately an 8σ perturbation! Or do we just mean the first object to form within a representative local volume, in which case considering a 3σ or 4σ perturbation may be sufficient.

5.3 How small are the first gas-rich protogalaxies?

- In the previous section, we saw how to construct the mass function of dark matter halos as a function of redshift. As z decreases, the mass function evolves in the direction of forming more massive halos. At some point, the characteristic mass of these halos will become large enough to exceed the mass scale required in order to induce gravitational collapse in the baryonic component of the Universe. Once this happens, gas will begin to fall into these halos, heralding the birth of the first dense gas clouds in the Universe.

- In order to quantify when this occurs, we need to be precise about what we mean by the "characteristic mass" of our halo mass function. This choice is somewhat arbitrary, but in studies of the formation of the first stars and galaxies, it is fairly common to take the mass corresponding to a 3σ density perturbation as a reasonable measure of the largest non-linear mass scale, which we will hereafter refer to as $M_{\rm NL}$.

- **SIMON: sketch evolution of $M_{\rm NL}$ with redshift**

- How large does $M_{\rm NL}$ need to be in order to induce collapse in the gas? From our discussion of structure formation in the linear regime, one obvious quantity that suggests itself is the Jeans mass

$$M_{\rm J} = \frac{1}{6}\pi^{-1/2} G^{-3/2} \frac{c_{\rm s}^3}{\rho_0^{1/2}} \tag{186}$$

- Since $c_{\rm s}^2 \propto T$, the Jeans mass scales with density and temperature as $M_{\rm J} \propto T^{3/2}\rho^{-1/2}$. At $z \gg 100$, the gas temperature is tightly coupled to the radiation temperature via Compton scattering, and hence evolves as $T \propto (1+z)$. Therefore, in this regime the Jeans mass evolves with redshift as $M_{\rm J} \propto (1+z)^{3/2}(1+z)^{-3/2} \sim$ constant.

- At $z < 100$, Compton scattering is no longer effective at maintaining $T_{\rm gas} = T_{\rm r}$, and the gas temperature evolves as $T \propto (1+z)^2$. In this regime, the Jeans mass evolves as $M_{\rm J} \propto (1+z)^{3/2}$.

- If we evaluate the Jeans mass in these two regimes, we find that at $z \gg 100$,

$$M_{\rm J} \simeq 1.4 \times 10^5 \left(\frac{\Omega_m h^2}{0.15}\right)^{-1/2} \, {\rm M}_\odot, \tag{187}$$

  while at $z \ll 100$,

$$M_{\rm J} \simeq 5.2 \times 10^3 \left(\frac{\Omega_m h^2}{0.15}\right)^{-1/2} \left(\frac{\Omega_b h^2}{0.026}\right)^{-3/5} \left(\frac{1+z}{10}\right)^{3/2} \, {\rm M}_\odot. \tag{188}$$

  (Note that the Ω_m and Ω_b terms are normalised here such that they are ~ 1 in our standard ΛCDM model).

- In the regime where $M_{\rm J}$ is constant, it is clear that halos with $M_{\rm NL} > M_{\rm J}$ will accumulate gas. However, in practice, at $z > 100$, $M_{\rm NL} \ll M_{\rm J}$. At lower redshifts, the Jeans mass becomes time-dependent and this presents us with a problem: which value of $M_{\rm J}$ do we compare with $M_{\rm NL}$? The current value? The value at turn-around? Or some other value?

- In practice, what we do is to look at an appropriately time-averaged form of the Jeans mass, known as the **filter mass**, $M_{\rm F}$. This is given by

$$M_{\rm F} = \frac{4\pi}{3}\rho_0 \left(\frac{\lambda_{\rm F}}{2}\right)^3, \tag{189}$$

  where the filter wavelength $\lambda_{\rm F}$ is given in the high redshift limit by

$$\lambda_{\rm F}^2 = \frac{3}{1+z}\int_z^\infty \lambda_{\rm J}^2 \left[1 - \left(\frac{1+z}{1+z'}\right)^{1/2}\right] {\rm d}z'. \tag{190}$$

- Evaluating this, we find that $M_{\rm F} = M_{\rm NL}$ at a redshift of around 30–40 (depending on the precise values chosen for our cosmological parameters, in particular σ_8). At this redshift, $M_{\rm F} \sim 3 \times 10^4$ M$_\odot$, around 50% larger than the instantaneous value of the Jeans mass.

- So far, we have been assuming that the gas starts at rest with respect to the dark matter. However, as we have discussed previously, this is now understood to be incorrect: the gas will be undergoing a bulk streaming motion with respect to the dark matter, owing to the residual effects of the strong gas-radiation coupling that was present at high redshifts. At recombination, the RMS velocity of this bulk flow is around $30\,\mathrm{km\,s^{-1}}$, but by $z \sim 30$, the expansion of the Universe has reduced this to around $1\,\mathrm{km\,s^{-1}}$. Nevertheless, this is still significantly larger than the sound speed at this epoch, and hence has a significant influence on the gravitational stability of the gas.

- Detailed modelling of the effects of these streaming motions shows that the net effect is to increase the minimum mass scale required for collapse by close to an order of magnitude. At $z \sim 30$, significant quantities of gas accumulate within the most massive dark matter halos only once $M_{\rm NL} > 2 \times 10^5\,\mathrm{M_\odot}$.

- So, does this mean that these objects are the sites where the first stars form? Not quite: this is a necessary condition for star formation, but not a sufficient condition.

- As gas begins to undergo gravitational collapse within one of these halos, it starts to heat up, owing to a combination of the effects of adiabatic compression and weak shocks. If the gas cannot dissipate any of this energy, then its rising temperature will lead to an increase in the pressure support that will eventually halt the collapse.

- We can use the virial theorem to estimate the mean temperature of the gas in the absence of dissipation. From the virial theorem, we know that at virialization, the kinetic energy of the halo is related to the potential energy by

$$|W_{\rm vir}| = 2T_{\rm vir}. \tag{191}$$

  For a spherical halo with an R^{-2} density profile, the gravitational potential energy can be written as

$$|W_{\rm vir}| = \frac{GM^2}{R_{\rm vir}}, \tag{192}$$

  where $R_{\rm vir}$ is the **virial radius** of the halo. (Note: real dark matter halos do not have R^{-2} density profiles, but changing from this to something more accurate only changes the value of $W_{\rm vir}$ by a small numerical factor of order unity.

- Within $R_{\rm vir}$, we know that for an idealized spherical perturbation, the overdensity with respect to the mean background density is $\Delta = 18\pi^2$. Hence, for a halo that virializes at a redshift $z_{\rm vir}$, we have

$$M = \frac{4\pi}{3}\Delta\rho_{\rm m,0}(1+z_{\rm vir})^3 R_{\rm vir}^3, \tag{193}$$

  where $\rho_{\rm m,0}$ is the present-day matter background density. Rearranging this, we find that

$$R_{\rm vir} = \left(\frac{3M}{4\pi\Delta\rho_{\rm m,0}}\right)^{1/3} \frac{1}{1+z_{\rm vir}}. \tag{194}$$

- The largest contribution to $T_{\rm vir}$ comes from the kinetic energy of the dark matter, but a fraction $\sim \Omega_{\rm b}/(\Omega_{\rm m}$ comes from the motion of the gas. If we assume that all of this energy is converted into heat, then the total thermal energy of the gas is given by

$$U_{\rm therm} = \frac{\Omega_{\rm b}}{\Omega_{\rm m}} \frac{|W_{\rm vir}|}{2}. \tag{195}$$

  (Note that we ignore the initial thermal energy of the gas, as this is typically negligible in comparison).

- We therefore have

$$U_{\rm therm} = \left(\frac{\pi}{6}\right)^{1/3} G M^{5/3} \Delta^{1/3} \frac{\Omega_{\rm b}}{\Omega_{\rm m}} \rho_{\rm m,0}^{1/3} (1+z_{\rm vir}). \tag{196}$$

  We can also write the thermal energy as

$$U_{\rm therm} = \frac{3}{2} N k T_{\rm vir}, \tag{197}$$

  where N is the total number of gas particles, and where we have assumed that the adiabatic index $\gamma = 5/3$. N is related to the halo mass via the expression

$$N = \frac{\Omega_{\rm b}}{\Omega_{\rm m}} \frac{M}{\mu m_{\rm p}}, \tag{198}$$

  where μ is the mean molecular weight, and so if we equate the two expressions above and do some rearrangement, we find that

$$T_{\rm vir} = \frac{2}{3} \left(\frac{\pi}{6}\right)^{1/3} \frac{G \mu m_{\rm p}}{k} M^{2/3} \Delta^{1/3} \rho_{\rm m,0}^{1/3} (1+z_{\rm vir}). \tag{199}$$

  Note that our final result is independent of $\Omega_{\rm b}$: increasing the baryon fraction increases the total amount of energy available as heat, but also increases the amount of gas that must be heated, and hence there is no change in $T_{\rm vir}$.

- Evaluating this for our standard cosmological parameters, we find that

$$T_{\rm vir} \simeq 400 \left(\frac{M}{10^5\ {\rm M}_\odot}\right)^{2/3} \left(\frac{1+z_{\rm vir}}{30}\right)\ {\rm K}. \tag{200}$$

  [Note that other definitions of $T_{\rm vir}$ exist in the cosmology literature that can easily differ by up to a factor of two from the value that we've derived here, depending on the assumptions made. $T_{\rm vir}$ should best be thought of as a rough estimate of the gas temperature prior to cooling, rather than a precise value]. Our $2 \times 10^5\ {\rm M}_\odot$ halo therefore has $T_{\rm vir} \sim 500\ {\rm K}$.

- After the gas has virialized, it will be able to undergo further gravitational collapse only if it can "cool" (i.e. dissipate thermal energy – note that its temperature may still increase, as long as the effective adiabatic index $\gamma_{\rm eff} < 4/3$, so that the Jeans mass decreases during the collapse).

- We know that the gas initially falling into our dark matter halo is predominantly atomic, with an ionization fraction of around 10^{-4} and an H_2 fraction of around 10^{-6}. At a temperature of 500 K, the gas is far too cold to cool via electronic line emission from atomic hydrogen or helium, and hence must rely on H_2 cooling.

- The mean hydrogen number density in a halo virializing at redshift z is given by

$$n_{\rm H} = \frac{\Delta\Omega_{\rm b}\rho_{\rm crit,0}}{\mu m_{\rm p}}(1+z_{\rm vir})^3 \sim 2\left(\frac{1+z_{\rm vir}}{30}\right)^3 \ {\rm cm}^{-3}. \tag{201}$$

  At this density and at a temperature $T \sim 500\,{\rm K}$, the H_2 cooling rate per unit volume is approximately $2\times10^{-25}\,{\rm erg\,s^{-1}\,cm^{-3}}$. For $x_{\rm H_2} = 10^{-6}$, we therefore obtain a cooling time

$$t_{\rm cool} = \frac{3}{2}\frac{nkT}{\Lambda_{\rm H_2}x_{\rm H_2}n^2} \sim 10^{18}\,{\rm s}. \tag{202}$$

  For comparison, the Hubble time at $z = 30$ is around $t_{\rm H} \sim 2\times10^{15}\,{\rm s}$.

- We therefore see that the amount of H_2 formed in the pre-galactic gas is not enough to produce effective cooling within the first protogalaxies. In order for the gas to cool, it therefore must form significantly more H_2 *in situ*.

- It is possible to construct a relatively simple model that captures the main features of the evolution of the H_2 fraction in the virialized gas. To begin, we assume that radiative recombination is the only process affecting the electron abundance in the gas. This allows us to write the rate of change of the electron number density as

$$\frac{{\rm d}n_{\rm e}}{{\rm d}t} = -k_{\rm rec}n_{\rm e}n_{\rm H^+}, \tag{203}$$

  where $n_{\rm e}$ is the number density of electrons, $n_{\rm H^+}$ is the number density of protons, and $k_{\rm rec}$ is the case B recombination coefficient.

- If we assume that ionized hydrogen is the only source of free electrons, implying that $n_{\rm e} = n_{\rm H^+}$, and that the temperature remains roughly constant during the evolution of the gas, then we can solve for the time evolution of the electron fraction:

$$x = \frac{x_0}{1+k_{\rm rec}ntx_0}, \tag{204}$$

  where $x \equiv n_{\rm e}/n$, n is the number density of hydrogen nuclei, and x_0 is our initial value of x.

- We next assume that all of the H_2 forming in the gas forms via the H^- pathway.[6] We also assume that the mutual neutralization reaction

$$H^- + H^+ \rightarrow H + H \tag{205}$$

[6] In practice, this is a reasonable approximation, as around 80–90% of the H_2 does indeed form via this route, with H_2^+ contributing only at the 10–20% level.

is the main process competing with associative detachment

$$\mathrm{H^- + H \rightarrow H_2 + e^-} \tag{206}$$

for the available $\mathrm{H^-}$ ions.

- These assumptions allow us to write the time evolution of the $\mathrm{H_2}$ fraction, $x_{\mathrm{H_2}} \equiv n_{\mathrm{H_2}}/n$, as

$$\frac{\mathrm{d}x_{\mathrm{H_2}}}{\mathrm{d}t} = k_{\mathrm{ra}} x n_{\mathrm{H}}\, p_{\mathrm{AD}}, \tag{207}$$

where k_{ra} is the rate coefficient of the radiative association reaction

$$\mathrm{H + e^- \rightarrow H^- + \gamma} \tag{208}$$

responsible for forming the $\mathrm{H^-}$ ions, and p_{AD} is the probability that any given $\mathrm{H^-}$ ion will be destroyed by associative detachment rather than by mutual neutralization.

- Given our assumptions, we can write the probability p_{AD} as

$$p_{\mathrm{AD}} = \frac{k_{\mathrm{ad}} n_{\mathrm{H}}}{k_{\mathrm{ad}} n_{\mathrm{H}} + k_{\mathrm{mn}} n_{\mathrm{H^+}}}, \tag{209}$$

where k_{ad} and k_{mn} are the rate coefficients for associative detachment and mutual neutralization, respectively. If $n_{\mathrm{e}} = n_{\mathrm{H^+}}$, $x \ll 1$ and $x_{\mathrm{H_2}} \ll 1$, this expression simplifies to

$$p_{\mathrm{AD}} = \left(1 + \frac{k_{\mathrm{mn}}}{k_{\mathrm{ad}}} x\right)^{-1}. \tag{210}$$

- Our expression for the time evolution of the $\mathrm{H_2}$ fraction therefore becomes

$$\frac{\mathrm{d}x_{\mathrm{H_2}}}{\mathrm{d}t} = k_{\mathrm{ra}} x n_{\mathrm{H}} \left(1 + \frac{k_{\mathrm{mn}}}{k_{\mathrm{ad}}} x\right)^{-1}. \tag{211}$$

- If the initial fractional ionization $x_0 \ll k_{\mathrm{ad}}/k_{\mathrm{mn}}$, then the term in parentheses is of order unity and this equation has the approximate solution

$$x_{\mathrm{H_2}} \simeq \frac{k_{\mathrm{ra}}}{k_{\mathrm{rec}}} \ln\left(1 + k_{\mathrm{rec}} n x_0 t\right), \tag{212}$$

$$= \frac{k_{\mathrm{ra}}}{k_{\mathrm{rec}}} \ln\left(1 + t/t_{\mathrm{rec}}\right), \tag{213}$$

where $t_{\mathrm{rec}} = 1/(k_{\mathrm{rec}} n x_0)$ is the recombination time. Since $k_{\mathrm{ad}} \sim 10^{-9}\,\mathrm{cm^3\,s^{-1}}$ and $k_{\mathrm{mn}} \sim 10^{-7}\,\mathrm{cm^3\,s^{-1}}$, this limiting solution applies provided that $x_0 \ll 10^{-2}$, which is clearly satisfied in this case.

- We see therefore that the growth of the $\mathrm{H_2}$ fraction is *logarithmic* in time. Most of the $\mathrm{H_2}$ forms within the first few recombination times, while at $t \gg t_{\mathrm{rec}}$, $\mathrm{H_2}$ formation slowly grinds to a halt.

- The main factor determining the final H_2 abundance is the ratio $k_{\rm ra}/k_{\rm rec}$, since for times of the order of a few recombination times, the logarithmic term in Equation 213 is of order unity. The two rate coefficients are given approximately by the simple power-law fits

$$k_{\rm ra} \simeq 1.83 \times 10^{-18} T^{0.88} \, {\rm cm}^3 \, {\rm s}^{-1}, \tag{214}$$

and

$$k_{\rm rec} = 1.88 \times 10^{-10} T^{-0.64} \, {\rm cm}^3 \, {\rm s}^{-1}. \tag{215}$$

(More accurate fits exist, with more complicated functional forms, but for the purposes of our current argument, these simple approximations suffice). The ratio of the two rate coefficients can therefore be written as

$$\frac{k_{\rm ra}}{k_{\rm rec}} \simeq 10^{-8} T^{1.52}. \tag{216}$$

- We see from this analysis that the amount of H_2 produced is a strong function of temperature. In practice, the largest fractional abundance that can be produced is a a few times 10^{-3}, as at very high temperatures, collisional dissociation of H_2

$$H_2 + H \rightarrow H + H + H, \tag{217}$$
$$H_2 + H_2 \rightarrow H + H + H_2, \tag{218}$$

etc., becomes effective and prevents x_{H_2} from becoming large. Therefore, gas-phase formation of H_2 via the H^- pathway never results in an H_2-dominated gas; atomic hydrogen always dominates.

- We have seen already that in our $2 \times 10^5 \, {\rm M}_\odot$ halo, the temperature of the gas will be around 500 K. Therefore, the fractional abundance of H_2 produced in the virialized gas after a few recombination times will typically be of the order of a few times 10^{-4}, i.e. a factor of 100 larger than the value in the pre-galactic gas. However, from our calculation above, we know that this is sufficient only to drop the cooling time in our example halo to $t_{\rm cool} \sim 10^{16}$ s, still a factor of a few larger than the Hubble time.

- We are therefore lead to the conclusion that the gas in this halo will not cool fast enough to undergo further gravitational collapse within a Hubble time. Since the survival time of a typical high-sigma dark matter halo is typically $\sim t_{\rm H}$, this implies that the gas in the halo will *never* form stars.

- We have seen that the amount of H_2 produced in the gas is a strongly increasing function of temperature. Moreover, the H_2 cooling function is also a steeply increasing function of temperature, meaning that the amount of H_2 that is required in order to cool the gas in less than a Hubble time *decreases* significantly with increasing T. Therefore, at any given redshift z, we can identify a critical temperature $T_{\rm crit}$, such that gas with $T > T_{\rm crit}$ will cool within a small fractionof a Hubble time, while gas with $T < T_{\rm crit}$ will not. Moreover, because the amount of H_2 formed in the gas and the amount required for efficient cooling are both relatively weak functions of z, the value of $T_{\rm crit}$ that we obtain is also relatively insensitive to redshift.

- In practice, we find that $T_{\rm crit} \sim 1000$ K over a wide range of redshifts. What does this then imply for the mass of the halo that is required? From our expression for the virial temperature, we see that $T_{\rm vir} \propto M^{2/3}$, and we know also that at $z = 30$, a halo with a virial temperature of 500 K has a mass of $2 \times 10^5\,{\rm M}_\odot$. This means that at this redshift, the first halos in which cooling is efficient will have masses $M \sim 2^{5/2} \times 10^5 \sim 6 \times 10^5\,{\rm M}_\odot$.

- These halos, with masses $\sim 10^6\,{\rm M}_\odot$, physical sizes of around 100 pc, and mean densities of a few particles per cm^3 are the sites in which the very first stars – the so-called Population III stars – form. In the next section we will examine the chain of events leading from the cooling of the virialized gas to their eventual formation.

6 Population III: formation and build-up of disc

- A useful way to think of the problem is to consider a collapsing blob of gas, where compressional heating, cooling and other heating processes all act to alter the energy of the gas. The rate of change of the energy per unit mass can be given as (Omukai 2000):

$$\frac{{\rm d}u}{{\rm d}t} = -p \frac{\rm d}{{\rm d}t}\left(\frac{1}{\rho}\right) - \Lambda_m + \Gamma_m \tag{219}$$

where the m subscripts on the heating and cooling rates denote that these have units of erg s^{-1} g^{-1}, and the pressure and energy per unit mass (or the "internal" energy) are given by $p = \frac{\rho kT}{\mu m_{\rm p}}$ and $u = \frac{1}{\gamma - 1}\frac{kT}{\mu m_{\rm p}}$. We can get the above expression into a more useful form by noting that

$$\frac{\rm d}{{\rm d}t}\left(\frac{1}{\rho}\right) = \frac{\rm d}{{\rm d}\rho}\rho^{-1}\frac{{\rm d}\rho}{{\rm d}t} \tag{220}$$

and that the evolution of the density with time is assumed to be controlled by gravitational collapse, such that

$$\frac{{\rm d}\rho}{{\rm d}t} \approx \frac{\rho}{t_{\rm ff}(\rho)} = \left[\frac{32{\rm G}}{3\pi}\right]^{1/2} \rho^{3/2} \tag{221}$$

yielding,

$$\frac{{\rm d}u}{{\rm d}t} = \frac{kT}{\mu m_{\rm p}}\left[\frac{32{\rm G}}{3\pi}\right]^{1/2} \rho^{1/2} - \Lambda_m + \Gamma_m. \tag{222}$$

Chemical rate equations typically work in volumetric units, rather than mass units, so we can convert the equation above to units of erg s^{-1} cm^{-3}, by multiplying through by ρ,

$$\frac{{\rm d}e}{{\rm d}t} = \frac{kT}{\mu m_{\rm p}}\left[\frac{32{\rm G}}{3\pi}\right]^{1/2} \rho^{3/2} - \Lambda_V + \Gamma_V. \tag{223}$$

where e is the energy density, and the V subscripts denote heating/cooling per unit volume.

- So how do atoms and molecules cool? Imagine a two level atom, with levels 0 (the ground) and an excited level 1. If no radiation field, then the number density of atoms in the excited level, is given by,

$$\frac{\mathrm{d}n_1}{\mathrm{d}t} = -(A_{10} + C_{10})n_1 + C_{01}n_0 \tag{224}$$

where A_{10} is the spontaneous emission rate (Einstein "A" co-eff), and C_{01} and C_{10} are the collisional excitation and de-excitation rates, respectively. The cooling rate per unit volume can then be given by,

$$\Lambda = C_{01}n_0E_{10}\ \frac{A_{10}}{A_{10} + C_{10}} \tag{225}$$

where E_{10} is the energy associated with the line transition between the two states. There a two limiting cases for this expression.

1. $A_{10} >> C_{10}$: the regime when radiative decay dominates. In this case the cooling is given by
$$\Lambda = C_{01}n_0E_{10} \tag{226}$$
Since $n_0 \equiv n$, and $C_{01} = q_{01}n$, where q_{01} is the collisional co-eff, then the above expression shows that $\Lambda \propto n^2$.
2. $A_{10} << C_{10}$: the regime when collisional decay dominates. In this case the cooling rate can be written as,
$$\Lambda = \frac{C_{01}}{C_{10}}A_{10}n_0E_{10} \tag{227}$$
If the system is evolving quickly enough, then one can set $\mathrm{d}n_1/\mathrm{d}t = 0$ in Eq.224 above, yielding
$$C_{10}n_1 = C_{01}n_0 \tag{228}$$
and the system can be said to be roughly in LTE, such that
$$\left(\frac{n_1}{n_0}\right)_{\mathrm{LTE}} = \frac{g_1}{g_0}e^{-E_{10}/kT} \tag{229}$$
and so
$$\Lambda = n_0\frac{g_1}{g_0}e^{-E_{10}/kT}A_{10}E_{10} \tag{230}$$
$$\Lambda = n_1A_{10}E_{10} \tag{231}$$
and so the cooling is proportional to n

the critical density at which this transition from non-LTE to LTE is given by $n_{\mathrm{cr}} = A_{10}/q_{10}$.

- In our first star-forming halos, T $\sim$ 1000 K and n $\sim$ 1cm^{-3}. Cooling by H_2 is still in the non-LTE regime, so given by $\Lambda_V = \Lambda_{H_2}x_{H_2}n^2$. Values Λ_{H_2} (note, erg s^{-1} cm^3!) can be found in Galli & Palla (1998, Fig 1A). We find that for a little above 1000K, the cooling rate is larger than the heating rate, so the gas cools as it collapses!

- DIFFERS FROM ISM COOLING!
- H_2 fraction continues to rise from its starting value of around 10^{-4}, but plateaus at round 10^{-3}, as the election fraction decreases due to recombination.
- However two features of H_2 start to kick in, that limit the temperature of the gas. One, the first accessible energy state is the J=2 rotational state at 512K. So H_2 cooling falls of exponentially at low T, and in practice can only cool the gas down to around few 100 K.
- Second, as the collapse proceeds, the gas density approaches the "critical density", at around $10^4 \mathrm{cm}^{-4}$, and the level populations start to come into LTE.
- The $p\mathrm{d}V$heating is now greater than the $\sim$ LTE cooling, Heating (puv) $\propto n^{3/2}$, while LTE cooling $\propto n$. As a result gas heats up as it collapses. However, even in LTE, H_2 cooling is still strongly dependent on T, so the rise in temperature with density is actually only gradual.
- At $10^4 \mathrm{cm}^{-4}$, we have a "loitering phase", where the collapse briefly halts. The temperature is around 250K. This sets a scale for the collapsing core. Jeans mass is 300 $\mathrm{M}_\odot$.
- Gas behaves like a polytrope as it collapses: $p = \rho kT/(\mu m_\mathrm{p}) = K\rho^{\gamma_{eff}}$, so the temperature and density are related via the effective polytropic index: $T \propto n^{\gamma_{eff}} - 1$.
- Draw both panels from Figure 3 in Yoshida et al. (2006) on the board, and label the points. Keep the figure, as we'll need to refer to it (or add to it) during the lecture.
- At around a density of $10^8 \mathrm{cm}^{-3}$, the conditions become favourable for H_2 formation via the following 3-body processes:

$$\mathrm{H} + \mathrm{H} + \mathrm{H} \rightarrow \mathrm{H}_2 + \mathrm{H} \tag{232}$$

and

$$\mathrm{H} + \mathrm{H} + \mathrm{H}_2 \rightarrow \mathrm{H}_2 + \mathrm{H}_2 \tag{233}$$

The rate for this reaction is incredibly uncertain, especially at low T (i.e. ≤ 500 K) and spans 2 orders of magnitude in the literature. It is also extremely difficult to measure in the lab. Glover (2008) rate is in the middle of the range (and is the one we adopt in our group), with $k_{3\mathrm{b}} = 7.7 \times 10^{-31} T^{-0.464}$ cm^6 s^{-1}, such that the rate of H_2 formation is (initially, when H_2 is sill small),

$$\frac{\mathrm{d}n_{\mathrm{H}_2}}{\mathrm{d}t} = k_{3\mathrm{b}} n_\mathrm{H}^3. \tag{234}$$

Taking $n_\mathrm{H} \approx n$, and $n_{\mathrm{H}_2} \sim 10^{-3} n$, we can calculate the H_2 formation time at $n = 10^8$ cm^{-3} and 1000K to be roughly,

$$t_{\mathrm{H}_2} \sim \frac{n_{\mathrm{H}_2}}{k_{3\mathrm{b}} n^3} \sim 10^5 \mathrm{yr} \tag{235}$$

If we compare this to the free-fall time at this density, we find $t_{\rm ff}(10^8 \mu m_{\rm p}) \sim 4400$ yr, which is considerably shorter. But remember that $\mathrm{d}n_{\rm H_2}/\mathrm{d}t \propto n^3$, while $t_{\rm ff} \propto n^{-1/2}$, so in practice, the H_2 fraction very quickly goes to 100%, over only a few decades in density evolution. Draw this on the plot from Yoshida.

- Rapid H_2 formation actually deposits energy in the gas, since the formation of each H_2 molecule releases 4.4eV (the molecular binding energy). As such, this chemical heating rate is given by,

$$\Gamma_{\rm 3b} = 4.4 \text{ eV} \frac{\mathrm{d}n_{\rm H_2}}{\mathrm{d}t} \tag{236}$$

  At around $n = 10^8$ cm^{-3}, $\Gamma_{\rm 3b} \sim 0.01\ \Gamma_{pdV}$, however by $n = 10^10$ cm^{-3}, $\Gamma_{\rm 3b} \sim 100\ \Gamma_{pdV}$, since $\Gamma_{\rm 3b} \propto n^3$, while $\Gamma_{pdV} \propto n^{3/2}$.

- So the formation of H_2 via the 3-body reactions actually stalls the collapse, at around $n = 10^9$ cm^{-3}, giving the reaction more "free-fall times" in which to complete. Once it is over, the gas is fully molecular. Note that because the formation rate is $\propto n^3$, as we more to higher densities, it becomes much faster than the dynamical time, so if H_2 is lost for any reason, it quickly reforms.

- The sudden appearance of H_2 via 3-body formation was actually suggested to cause a chemothermal instability, which could lead to $t_{\rm cool} < t_{\rm ff}$ for a very short period. This lead to suggestion that it could lead to fragmentation. Yoshida performed a stability analysis of this phenomenon, showing that some of the gas does indeed go unstable, however the amount of gas in this phase, and the duration of the instability is not long enough to promote fragmentation: only a single collapsing centre continues.

- At higher densities, above $n = 10^{10}$ cm^{-3}, the cooling by H_2 starts to become optically thick, and the efficiency at which H_2 powers the collapse decreases. How de we treat that? Optically thick line radiative transfer is notoriously difficult, especially in 3D. Ripamonti & Abel (2004) found an empirical fit of the form

$$\Lambda_{\rm H_2,thick}(T) = \Lambda_{\rm H_2,thin}(T) \min\{1, (n/n_0)^{-\beta}\} \tag{237}$$

  with $n = 8 \times 10^9$ cm^{-3} and $\beta = 0.45$.

- A better way to treat this in numerical calculations is to adopt an "escape probabilty" formalism. The net cooling rate is given by,

$$\Lambda_{\rm H_2,thick}(T) = \sum_{u,l} h\nu_{ul} \beta_{\rm esc,ul} A_{ul} n_u \tag{238}$$

  where n_u is the population density of the H_2 in the upper energy level u, A_{ul} is the Einstein coefficient for spontaneous trasnition, $\beta_{\rm esc,ul}$ is the escape probability for a photon with frequency ν_{ul} to escape from the parcel of gas in question and $h\nu_{ul}$ is the energy difference between the two levels.

  The escape probability is related to the optical depth via (see Stahler & Palla),

$$\beta_{\rm esc,ul} = \frac{1 - \exp(-\tau_{\rm ul})}{\tau_{\rm ul}}, \tag{239}$$

where we approximate $\tau_{\rm ul}$ as

$$\tau_{\rm ul} = \alpha_{\rm ul} L_{\rm s} \tag{240}$$

where $\alpha_{\rm ul}$ is the line absorption coefficient and $L_{\rm s}$ is the Soloblev length. In the classical, one-dimensional spherically symmetric case, the Sobolev length is given by

$$L_{\rm s} = \frac{v_{\rm th}}{|{\rm d}v_{\rm r}/{\rm d}r|}, \tag{241}$$

where $v_{\rm th}$ is the thermal velocity, and ${\rm d}v_{\rm r}/{\rm d}r$ is the radial velocity gradient. If the velocity dispersion of the gas is very small, then $L_{\rm s}$ can become very large, much larger than the size of the collapsing core. To ensure that the H_2 cooling rate is not reduced to an artificially low value, it makes sense to use the smallest of the Soloblev length and the local Jeans length, $L_{\rm J}$ in the expression for τ_{ul}.

Since the line absorption coefficient $\alpha_{\rm ul}$ is linearly proportional to the number density of H_2, we can write $\tau_{\rm ul}$ as

$$\tau_{\rm ul} = \left(\frac{\alpha_{\rm ul}}{n_{\rm H_2}} \right) N_{\rm H_2,eff}, \tag{242}$$

where $N_{\rm H_2,eff} \equiv n_{\rm H_2} L_{\rm s}$ is an effective H_2 column density, and where $\alpha_{\rm ul}/n_{\rm H_2}$ is a function only of temperature. We therefore tabulate the cooling rate per H_2 molecule in the optically thick limit as a function of two parameters: the gas temperature T and the effective H_2 column density $N_{\rm H_2,eff}$, and compute cooling rates during the simulations by interpolation from a pre-generated look-up table.

- Although H_2 line-cooling is gradually shut-off, another cooling mechanism comes into play: collision induced emission. At very high densities, the collisions between pairs of molecules can temporarily form a "super molecule", which can have a dipole. There is small probability that this can result in a dipole-transition, and with enough collisions, result in cooling via continuum emission. Around $n = 10^{13}$ cm^{-3}, this starts to become the dominant coolant – and just in time, as standard H_2 line-cooling is now very inefficient. However, at around $n = 10^{15}$ cm^{-3}, this too starts to become optically thick.

- Finally, the gas one possible cooling channel left: it can dissociate H_2. Each 4.4eV per molecule that heated the gas during the 3-body formation can now be reclaimed. Acts as a sink for the $p{\rm d}V$ heating, helping to keep the collapse going.

- When the H_2 runs out, a hydrostatic core is finally born. The mass of this object is around few 0.01 $M_\odot$ with a radius of around 0.04 AU (around 10 $R_\odot$).

- Draw the radial plots... Greif?

- So how quickly does this star grow? A simple estimate can be given by considering the mass of the collapsing core and the free-fall time,

$$\frac{{\rm d}M_*}{{\rm d}t} \sim \frac{m_J}{t_{\rm ff}(n)} \sim \frac{c_{\rm s}^3}{G}. \tag{243}$$

  For our initial collapse properties in the minihalo, this yields a rate of roughly 10^{-3} $M_\odot$ yr^{-1}. Over two orders of magnitude greater than present-day star formation.

- A more accurate picture of the accretion onto the star can be obtained by looking at the radial profiles of the density and radial velocity, at the point at which the central hydrostatic core forms. The temporal evolution of the accretion rate can be given by,

$$\frac{\mathrm{d}m(r)}{\mathrm{d}t} = 4\pi\, r^2\, \rho(r)\, v_{\mathrm{r}}(r) \tag{244}$$

where $m(r)$ is the mass of a shell at radius r, and $\rho(r)$ and $v_{\mathrm{r}}(r)$ are, respectively, the density and radial velocity in the shell. **Draw this on the board. Take from the SOM of Clark et al. 2011**. Using the fact that mass enclosed at given shell R is,

$$m_{\mathrm{enc}}(R) = 4\pi \int_0^R r^2\, \rho(r) \mathrm{d}r \tag{245}$$

one can relate the mass accretion rate to the mass that has fallen to the centre (i.e. the star), yielding an expression for $\dot{m}(m_*)$. **Draw this too**.

- However accretion doesn't proceed directly onto star. Conservation of angular moment results in the build-up of a disc. How quickly can a disc drain onto the star? Consider a disc comprising circular shells at radius r, within a disc with mean surface density $\Sigma = M_{\mathrm{disc}}/\pi R_{\mathrm{disc}}^2$. The mass of a shell is then given by,

$$\mathrm{d}M \simeq 2\pi\, r\, \mathrm{d}r\, \Sigma \tag{246}$$

and the flow through the disc is given by,

$$\frac{\mathrm{d}M}{\mathrm{d}t} \simeq 2\pi\, r\, \frac{\mathrm{d}r}{\mathrm{d}t}\, \Sigma \simeq 2\pi\, r\, v_r\, \Sigma. \tag{247}$$

If there is no viscosity in the disc, then in a steady state disc, v_r and there would be not accretion through the shell.

Can characterise rv_r as the "kinematic viscosity", ν. Normally assumed to be a local phenomena, such as turbulent mixing, but can also be produced via-global gravitational torques and the associated Reynold's stresses that they create in the gas. Shakura & Sunyaev (1973) parameterised the viscosity in terms of α, with $\nu = \alpha H_{\mathrm{p}} c_{\mathrm{s}}$, where H_{p} is the pressure scale-height, a measure of the disc's thickness.

The rate of mass flow through the disc and on to the protostar, can then be estimated by,

$$\frac{\mathrm{d}M}{\mathrm{d}t} \simeq 2\pi\, \alpha H_{\mathrm{p}} c_{\mathrm{s}} \Sigma. \tag{248}$$

A more careful analysis will yield a pre-factor of 3, rather then 2, but then this equation is strictly only valid for thin-discs, so our analysis is, in any case, rather rough. Our recent work (Clark et al. 2011, Science), yields the following numbers: $\Sigma = 5 \times 10^3$ g cm^{-2} ($n = 5 \times 10^{13}$ cm^{-3}), T = 1500 K so $c_{\mathrm{s}} = 2.4$ km/s, $H_{\mathrm{p}} = 2$ AU ($\sim \Sigma/\rho$). The maximum value of α that one can expect is in the range 0.1 to 1, for strongly self-gravitating discs. Together this yields an accretion rate *through* the disc of around a few 10^{-4} M$_\odot$ yr^{-1}. Disc is unable to process infalling gas, and grows larger.

- If disc grows sufficiently large, it can become gravitationally unstable, and fragment to form new stars. A thorough treatment of this was presented by Toomré, but a simplified treatment can be given as follows. Consider an element of the disc with surface density Σ and radius R. The gravitational acceleration of this element is given by $a_{\rm g} \simeq G\Sigma$. The supporting shear term is given by $a_{\rm s} = R\Omega$, and the supporting pressure is given by roughly $p \simeq \Sigma c_{\rm s}^2$, giving $a_{\rm p} = c_{\rm s}^2/R$. The total repulsive force is therefore $a_{\rm s} + a_{\rm p} = R\Omega + c_{\rm s}^2/R$. This diverges at both large and small R, but has a minimum at $2c_{\rm s}\Omega$. If this minimum is smaller than the gravitational term, then this element of the disc is unstable: $2c_{\rm s}\Omega < G\Sigma$. This then sets the stability $Q = 2c_{\rm s}\Omega/G\Sigma > 1$ for a stable disc. Proper analysis yields $Q = \kappa c_{\rm s}/\pi G\Sigma$, where κ is the epicyclic frequency. For roughly Keplerian discs, $\kappa \simeq \Omega$.

- We found that the discs around Pop III stars do indeed fragment, as $Q < 1$. Accretion luminosity heating can help stabilise the inner disc, but the fragmentation occurs futher out (20-30 AU instead of around 10 AU). Previous estimates predicted gravitationally stable discs, as they assumed disc would be atomic (and therefore much hotter). Numerical simulations show that this is not the case, and the disc is H_2 rich.

- Such self-gravitating discs typically display prominent spiral arms. Draw picture. In the spirals, the density can be much larger than the ambient density in the disc (factor of 10 or so), and it is in these arms that the fragmentation occurs. Find $n = 1 \times 10^{14}$ cm^{-3} and $T \sim 1500$ K. The Jeans mass is 0.05 $M_\odot$: in the sub-stellar regime! If fragmentation results in many objects, then we could get an ejection via dynamical encounters – some of these objects could have be ejected from the mini-halo when they were very low mass. If they have $M_* < 0.8$ $M_\odot$, then they could still be on the MS today!

- Complication 1: dark-matter annihilation could provide a source of heating during the collapse, if the DM is made up of WIMPS. Freese et al. (2008) and Spoylar et al. (2009) proposed that this could halt the collapse entirely, forming a "Dark Star" - a failed star, powered by DM-annihilation. Recently, Smith et al. (2012; our group) has shown that this doesn't work, as the collisional dissociation of H_2 can keep the collapse going until the DM-heating phase is over. However, depending on the mass of the DM candidate, it seems that much of the fragmentation can be reduced. However the whole mechanism relies on collapse occurring directly on around the central DM cusp.

- Complication 2: our group has done a lot of work showing that the small-scale dynamo can amplify a small seed field during the collapse of the gas in the minihalo, with the suggestion that the magnetic energy will saturate close to equipartition once the gas reaches the density of the disc. However the field is initially suspected to be incoherent, and so its dynamical effect is still unknown. This remains a hot topic in Pop III star formation research.

7 Population III: growth of the protostar and feedback

- The goal of this section will be to describe the evolution of the central protostar as it accretes from the natal core. In general terms, the energy released by the gas as it shocks with the protostellar surface is given by,

$$L_* \simeq \frac{G\,M_*\,\dot{m}}{R_*} \tag{249}$$

- The temperature, radius and luminosity of the star are related by

$$L_* = 4\pi\,R_*^2\,\sigma_{\rm SB} T_*^4 \tag{250}$$

  where $\sigma_{\rm SB}$ is the Stefan-Boltzmann constant.

- The aim of protostellar evolution theory is to solve for the protostar's structure while it is still being assembled, to self-consistently solve for R_* and T_* (or for $R_{\rm ph}$ and $T_{\rm ph}$). Clearly this is difficult. Normally requires that we split the problem into two pieces: 1) study how the protostar grows under constant accretion 2) deduce how the resulting protostar can affect the accretion flow.

- The accretion onto a protostar is associated with two main timescales, and we start by describing these.

  1. The *accretion timescale*, the time over which the protostar increases its mass:

$$t_{\rm acc} = \frac{m_*}{\dot{m}_*} \tag{251}$$

  2. The *Kelvin-Helmholtz timescale*, the time over which the gravitational energy of a core (or in this case a protostar) is radiated away:

$$t_{\rm KH} \equiv \frac{GM_*^2}{R_* L_*} \tag{252}$$

  It is clear from the above discussion of the accretion rate that both these timescales are changing with time. As such we shall see that protostellar evolution is divided into distinct phases that depend on which of the above timescales dominates (i.e., is the shortest).

- Introduce the work of Stahler et al (1980; 1981, 1986), and Hosokawa & Omukai (2010).

- The basic structure that is found from solving the stellar structure equations can be summarised by drawing Figure 1 from SPS86.

- Draw Fig. 6+17 from HO12. General outcome of the numerical studies is that the protostellar evolution is characterised by the follow phases:

1. *The Adiabatic Accretion Phase* - The gas in the core is initially extremely optically thick, resulting in a low luminosity. As such, $t_{\rm acc} >> t_{\rm KH}$, and the core continues to grow as the new material is deposited at its surface in the accretion shock. Note that as the post-shock entropy increases over time due to the increasing strength of the accretion shock (which is a function of m_*), the core develops an off-centre distribution of entropy and temperature.

   The gas around the core remains optically thick, in what is termed a "radiative precursor" (see diagram). The opacity is provided by H^-, which is created in the shock. The result is an adiabatically evolving core surrounding by a an optically thick region (the radiative precursor). In contrast to present-day protostars, with their lower accretion rates, the Pop III stars remain radiatively supported during this phase.

   SPS86 find the following relations during the adiabatic phase:

$$R_* = 48.1 R_\odot \left(\frac{M_*}{M_\odot}\right)^{0.27} \left(\frac{\dot{M}}{4.41 \times 10^{-3} M_\odot}\right)^{0.41} \tag{253}$$

$$R_{\rm p} = 66.8 R_\odot \left(\frac{M_*}{M_\odot}\right)^{0.27} \left(\frac{\dot{M}}{4.41 \times 10^{-3} M_\odot}\right)^{0.41} \tag{254}$$

$$T_{\rm p} = 5170\,{\rm K} \left(\frac{M_*}{M_\odot}\right)^{0.044} \left(\frac{\dot{M}}{4.41 \times 10^{-3} M_\odot}\right)^{-0.055} \tag{255}$$

2. *The Swelling Phase* - As the core gradually contracts, its temperature slowly increases. Opacity is a strongly decreasing function of temperature in the core, and eventually the gas becomes optically thin enough to allow a distribution of the internal entropy in the core. As the temperature is highest in the centre, this change in opacity begins in the deep layers of the core and works its way to the surface in a "luminosity wave". Once the core reaches about 5 $M_\odot$, this distribution of entropy is fast enough to cause the protostar to rapidly expand (remember that a star has a negative specific heat - core contracts, so outer layers must expand). As the opacity at a given temperature is lower in the primordial case than in present-day protostars, this phase can occur slightly earlier.

3. *The Kelvin-Helmholtz Contraction Phase* Once the luminosity wave reaches the surface, the energy escapes the protostar - it signifies the point at which all parts of the star are now able to lose heat. Now $t_{\rm KH}$ is slightly less than $t_{\rm acc}$. The protostar starts to contract, and the luminosity from this contraction now dominates over the accretion luminosity.

4. *The Arrival at the Main Sequence* - Once the temperature in the contracting star reaches $\simeq 10^8$ K, the star is finally able to form enough C via He burning, that it can start a CN cycle, and achieve fusion support. For solar metallicity stars, which already have C and N, this process occurs at much lower temperatures ($\simeq 10^7$ K), and hence Pop III stars are smaller than their present-day counterparts.

- Except for a short period during the swelling phase, the radiative precursor remains intact. Due to the strong shock at the core surface, the gas becomes ionised (collisional), and sufficient H^{-1} forms to provide a high (bound-free) opacity. As the H^{-1} abundance is a strong function of temperature, the precursor becomes more optically thick as it heats up. As such, the entire evolution of the star is determined by two things: how fast the entropy comes in, and how quickly it can leave the precursor. This keeps the precursor at a roughly constant temperature of between $6000 - 7000$K. Thus, in the spherically symmetric model, the star has a roughly constant photospheric temperature.

- Tan & McKee (2004) and McKee & Tan (2008) looked at a simple model for accretion via a disc. In this case, they found that as the accretion occurs further out, the conditions around the shock are optically thin, and the surrounding gas will see much more the bare stellar surface, which has a much higher temperature than the precursor in the spherically symmetric case. Actually, they assume that most of the H_2 in the disc has been dissociated in the accretion shock that occurs at $r > R_{\rm disc}$, and so assume that the disc is atomic and thus small. Neglects the rapid 3-body H_2 formation rate!

7.0.1 The final fate of Pop III stars

- When do Pop III stars stop accreting? What determines their final mass? There are many possible mechanisms for terminating the accretion onto a Pop III star, including death (SN), HII region expansion, dissociation of H_2 via Lyman-Werner radiation, and radiation pressure. Although the current debate tends to focus on the effects of ionising radiation from Pop III stars, it is likely that all these processes play a role. We discuss them each in turn, highlighting their possible contribution.

- Perhaps the simplest to understand is SN feedback. The life of a massive star is only $\sim$ 2 Myr for $M_* > 100 M_\odot$. The accretion timescale in the baryons in the minihalo is around 10 Myr (or $t_{\rm ff}$ at around a number density of 1 cm^{-3}), so if nothing stops the star from accreting, then it could die before the reservoir is drained. The expected outcomes for Pop III stars of varying progenitor mass are as follows

  1. 15 - 40 $M_\odot$ $\rightarrow$ Core-collapse SN ($\sim 10^{51}$ erg)
  2. 40-140 $M_\odot$ $\rightarrow$ Collapse to BH, no remnant
  3. 140-260 $M_\odot$ $\rightarrow$ PI SN (10^{51} - 10^{52} erg)
  4. $>$ 260 $M_\odot$ $\rightarrow$ Collapse to BH!

- The binding energy of the first star-forming minihalos is around 10^{50} erg, and so even in the event of no other feedback process, this could clear the halo of baryons, preventing accretion onto siblings. Suggests a rough upper limit to Pop III stars of a few 100 $M_\odot$.

- Another option is that feedback from the young star can dissociate the H_2 in the collapsing gas, thus removing its coolant. This could halt the collapse, as the effective EOS of the gas is now roughly adiabatic. According to McKee & Tan (2008), this

shouldn't be a problem if the infall is supersonic onto a star. However this assumes that mass is dominated by central region (the star+disk) and not the envelope. It currently remains unclear how this will proceed. Note that the gas will heat up until it can cool effectively via (LTE) Lyman Alpha emission.

- H2 can be dissociated by Lyman-Werner photons, i.e. those with energies in the range 11.15 to 13.6 eV. Note that not all photons will result in dissociation, as not all will be absorbed by the Lyman-Werner lines. Those that are can they decay to lower vibrational levels an potentially dissociate the molecule (for a discussion, see Glover & Brand 2001). While a promising mechanism on largest scale within the halo, it is not clear that this process can shut-off accretion close to the star. As such, by themselves they could probably prevent the entire minihalo collapsing, but are unlikely to limit the mass to below a few 100 $M_\odot$. We'll discuss the LW photons in more detail later.

- The more promising mechanism for shutting off accretion is via the expansion of an HII region around the central star. Massive stars can release a substantial number of ionising photons (i.e. those with $h\nu > 13.6$ eV), especially once they have reached the ZAMS. The basic physics of HII regions is fairly well established, however their interaction with the complex accretion geometry that arises during star formation has been only recently been studied. We will first summarise the basics of HII region formation, before discussing the results from the recent numerical studies.

- We start with a neutral gas with number density n_0 and ionise it, such that it has a number density n_i or ions and n_e of electrons. If the gas is HI, then $n_i = n_e$. The rate of recombinations per unit volume is then given by $\alpha_{\rm B} n_i n_e$, where $\alpha_{\rm B}$ is recombination coefficient for "case B" recombinations, which neglects those recombinations directly to the ground state. These other recombinations are assumed to emit photons that are absorbed elsewhere in the HII region – the so-called "on-the-spot" approximation (if the mean-free-path of these photons is smaller than the HII region, then this is justified). The coefficient $\alpha_{\rm B}$ for a gas of pure H has a value 3×10^{-13} cm^3 s^{-1}. If the gas is fully ionised, then we can write this as simply $\alpha_{\rm B} n_0^2$. Now imagine we have a source Q_* of ionising photons (units of per second), that ionise the gas in spherical region with density n_0. The number if ionising photons arriving at a radius R, is given by

$$4\pi\, r^2\, J(R) = Q_* - \int_0^{4\pi} \int_0^R r^2 \alpha_{\rm B} n_0^2 \, {\rm d}r {\rm d}\Omega \tag{256}$$

where $J(R)$ is the flux of ionising photons passing through the surface bounded by r, and Ω is the angular dependancy. The equation simply expresses that the number of photons reaching r is number leaving the star minus those that are required to balance recombinations. At some radius, the integral on the RHS of the above expression will equal Q_*, and the flux of ionising photons will drop to zero. This radius, named the Strögren radius, is given by

$$R_{\rm S} = \left(\frac{3 Q_*}{4\pi \alpha_{\rm B} n_0^2} \right)^{1/3} \tag{257}$$

And denotes the maximum initial radius of the HII region that a star can maintain. However it takes time for this region of ionised gas to develop. For the ionisation front to move a distance dr will require a certain number of photons. In fact, $J(r)\mathrm{d}t$ photons are required to fully ionise $n_0\mathrm{d}r$ amount of gas, such that $J(r)\mathrm{d}t = n_0\mathrm{d}r$. So the ionisation front propagates at a speed $\mathrm{d}r/\mathrm{d}t = J(r)/n_0$. We can now use this to write,

$$\frac{\mathrm{d}r}{\mathrm{d}t} = \frac{J(r)}{n_0} = \frac{Q_*}{4\pi r^2 n_0} - \frac{1}{3} r \alpha_{\mathrm{B}} n_0 \tag{258}$$

This equation describes initial the advance of the i-front into the neutral medium. In this phase, the i-front travels quickly (initially $>> c_s$), as the photons eat their way into the surrounding gas. This type of front is termed an "R" front. Once the total number of photons and recombinations are equal, the Strömgren sphere stage is reached. However, at this point, the ionisation region will continue to expand, since the gas in the region is substantially hotter ($\sim 10000K$) than the surroundings, while still have roughly the same density. The result is a pressure driven shock that drives further expansion of the i-front – this time, termed "D"-front. This phase finally comes to end when the pressure in the HII region equals that of the surroundings.

- Taking the spectrum of a star to be that of a black-body, the rate of ionising photons is given by,

$$Q_* = \frac{\pi L_*}{\sigma_{\mathrm{SB}} T_{\mathrm{eff}}^4} \int_{\nu_{\mathrm{min}}}^{\infty} \frac{B(T_*)_\nu}{h\nu} \mathrm{d}\nu \tag{259}$$

- The above expression for the Strögren sphere is found to work well when the density is uniform. But what about when we have a steep density profile such as those found in star formation? If $n_0 \propto r^{-\omega}$, we find that for $\omega > 1.5$, solution diverges: the integral becomes:

$$Q_* \propto (3 - 2\omega)^{-1} r^{3-2\omega} \tag{260}$$

In reality what this means is that HII runs down the density profile unimpeded, remaining R-type as it does so, until it encounters gas with a shallower density gradient, and can start to enter an R_{s}-like phase. Typically Pop III star formation has $\omega = 2.2$ initially, which evolves to $\omega = 1.5$ as the star grows in mass (the typical solution for matter falling onto a point mass).

- However very close to the star, the density profile is much flatter, and so the HII region must first "break out" of this inner region before engulfing the halo. Taking $n_0 = 10^{13}$ cm^{-3} in a region of 10 AU (i.e. the typical properties of the disc in Clark et al. 2011), we would need a source of $Q_* = 10^{56}$ s^{-1} to break out. Can we get such high numbers? The answer is actually no, and in fact the break out of the i-front takes around 10^4 years.

- Hosokawa et al result present 2D simulations of the collapse of a rotating Pop III star-forming cloud onto a central star. They follow the ionising radiation using a raytracer and diffuse radiation using FLD (no on-the-spot approximation). The properties of the central source are calculated self-consistently using the set-up from Hosokawa &

Omuaki (2010). However they adopt R_* rather than R_p when calculating the stellar spectrum. Focuses on scales > 10 AU.

- The basic result of H-etal12 is that the accretion onto the central star is terminated when the star reaches 43 $\mathrm{M}_\odot$. Sketch their Fig 1 and 2.

- Even at late times, $t_\mathrm{acc} > t_\mathrm{ion}$ in the disc, and only as the star approach 43 $\mathrm{M}_\odot$ can it finally shut of the disc accretion on scales of < 100 AU. The erosion of the larger disc is going to take significantly longer.

- Stacy et al. 2012: lower resolution, but looks at larger scales and in 3D. Uses a simplified prescription for the PMS/MS model, roughly based on Hosokawa et al 2010, and again adopting the stellar radius as the photosphere, rather than the radiative precursor. They find that the Lyman-Werner radiation plays a more dominant role, but this is likely just because they have a lower density medium. In fact, they follow scales (and thus densities) in which 3-body H2 formation is unimportant. However their results suggest that once the HII region has broken apart the cloud, the Lyman-Werner radiation will be able to sterilise much of the halo.

- Note that Lyman-Werner has to keep pace with 3BH2 formation, which goes as n^3 (at least for $n > 10^8$ cm$-$3), while ionisation has to keep up with H_2 formation, which goes as n^2.

- Do the $T(\rho)$ plot from Stacy et al, and show the various phases of the gas.

- Breakout of HII region? Can it clear the Halo? Whalen looked at 150+ $\mathrm{M}_\odot$ sources (ie. those with ionising photon counts greater than $> 10^{50}$ /s), and found that star would typically clear the min halo, in all but the most massive halos. However if Pop III stars have lower masses, then these arguments might not hold. The Hosokawa star has only a few an ionising photon count of 10^{49} s^{-1}, much less than those studied by Whalen.

- Finally, Hosokawa shows that final mass can depend on the rotation in the halo. Lower rotation, leads to higher densities in the region around the star, which delays the breakout. The accretion rates onto the star are also higher, but the extra radiation is not enough to compensate for the faster growth of the star plus higher density environment (note that the environment has an n^2 effect).

- Can the pressure exerted by photons on the gas not blow the accretion flow away? The Eddington Luminosity defines the point at which radiation pressure is balanced by gravity. First consider the case of hydrostatic equilibrium:

$$\frac{\mathrm{d}v_\mathrm{r}}{\mathrm{d}t} = -\frac{\nabla p}{\rho} - \nabla\phi = 0 \tag{261}$$

In the case where the dominant source of pressure is from radiation, then

$$-\frac{\nabla p}{\rho} = \frac{\kappa F_\mathrm{rad}}{c} \tag{262}$$

where $F_{\rm rad}$ is the radiation flux and κ is the opacity, given by $\sigma_{\rm T}/m_{\rm p}$, and $\sigma_{\rm T}$ is the Thompson scattering cross-section. The luminosity passing through a surface S is then given by,

$$L = \int_S \mathbf{F}_{\rm rad} \cdot {\rm d}\mathbf{S} = \int_S \frac{c}{\kappa} \nabla\phi \cdot {\rm d}\mathbf{S} \tag{263}$$

If κ is the same everywhere within the shell, then the Eddington luminosity is given by,

$$L_{\rm Edd} = \frac{c}{\kappa} \int_S \nabla^2\phi {\rm d}V = \frac{4\pi G c}{\kappa} \int_S \rho {\rm d}V = \frac{4\pi G M c}{\kappa} \tag{264}$$

Note that M here is the mass enclosed by a shell at r, so the formula can be used for any point in the spherical geometry. A shell emitting a luminosity higher than $L_{\rm Edd}$ will be able to repel a flow that is free-falling towards it. The formulation as it stands is for a pure H plasma, considering only scattering with free-elections, however the basic picture still holds, and one can replace the scattering opacity with the that of the species in question. For present-day star formation, the debate typically focuses around the effects of dust.

- For Pop III stars, there is obviously no dust, so the most commonly discussed option for radiation pressure is that from Lyman-α photons and from the electron scattering. These require that the gas has already been either dissociated or ionised, and so can be thought of as additional effects on top of those already discussed. MT08 suggest that these effects can be almost as strong as the gas pressure in the two scenarios, and so they should help to cut off accretion onto the Pop III star.

- So the current state-of-the-art would suggest that Pop III stars are massive (~ 40 M$_\odot$), but not typically the monsters that would fuel PI-SN. However it would seem that they are able to clear their halo of baryons within their lifetimes.

- Starvation induced fragmentation? Can we grow such massive stars? Note that if you have a distributed cluster, you avoid the very high column densities that were seen in the Hosokawa et al. study, that blocked the Lyman-Werner radiation.

7.1 The Pop III.2 channel

- Up until now, we have explored the Pop III(.1) formation channel, but there is also a Pop III.2!

- Occurs in halos that have been exposed to the radiation field of the very first (.1) stars. Increased ionisation produces free electrons that help catalyse H2. Allows the gas to cool down further than in Pop III.1 case.

- x(D) = 2.6e-5 (again, relative to H).

- Reactions:

$$\mathrm{H_2 + D^+ \rightarrow HD + H^+} \tag{265}$$

$$\mathrm{HD + H^+ \rightarrow H_2 + D^+} \tag{266}$$

- Reaction 1 is exothermic but reaction 2 is endothermic by the time the gas reaches 462K. As a result, chemical fractionation occurs at low temperatures, enhancing the ratio of $x(\mathrm{HD})/x(\mathrm{H_2})$ to $x(\mathrm{D})/x(\mathrm{H})$. Equilibrium fraction value : $x(\mathrm{HD})/x(\mathrm{H_2}) = x(\mathrm{D})/x(\mathrm{H})e^{(462/T)}$. At 200K, this is already a factor of 10.
- Plot the cooling plot from Glover 2008 (the conference proceedings). Show that the cooling becomes comparable to the H2 cooling below 200K or so.
- cooling is so efficient that the gas gets cold enough to make use of HD cooling, allowing it to cool down to the CMB temp ($T_{\mathrm{CMB}} = 2.728(1+z)$).
- Plot the $T(\rho)$ diagram for Pop III.2
- Note the Jeans mass (~ 30 $\mathrm{M_\odot}$) and the star formation rate (few 10^{-5} $\mathrm{M_\odot\,yr^{-1}}$).
- Although colder, an thus smaller initial Jeans masses, we tend to find less fragmentation! 1) lack of structure due to stiff EOS once HD/H2 cooling goes to LTE 2) Lower accretion rate! However, it should be stressed that the evolution of the protostellar disc around a Pop III.2 star has never been properly followed in a full cosmological simulation.
- Hosokawa et al 2012b performed a similar simulation to their Pop III.1 case, and found again that the feedback (ionisation) can halt the accretion when the central object has a mass of around 15 $\mathrm{M_\odot}$. Again, if the rotation is slower, then the star is larger (due to the higher densities everywhere around the star). They conclude that ALL Pop III stars (both .1 and .2) should be in the lower-mass range (10 - 50 $\mathrm{M_\odot}$).
- Note that this channel can also be activated when the halo has a greater mass than in the "typical" Pop III(.1) case, as the virial shock can also collisionally ionise enough H to increase the election fraction.
- Not actually a good terminology, since the gas is still primordial, and there is no real clear starting point for Pop III.2, since the importance of HD cooling is a continuous function of the halo size.

8 Atomic cooling halos

- In yet larger halos, the virial temperature is large enough to excite Lyman-alpha cooling. These are termed the "first galaxies", as they don't require H2 to form, and can survive fairly strong background radiation fields - including those from Ionising stars. First objects that can harbour, and retain, star formation. Masses around 5e7 Msun, Rvir $\sim$ kpc. Often referred to as the 'first galaxies'.
- Draw something like Fig. 1 in Greif et al 2008 to show how common these halos are with redshift. Only form in significant numbers around z = 10, but the first objects appear around z = 15 (roughly 140 Myr later after the formation of Pop III star-forming minihalos).

- Star formation can proceed via two channels in such halos 1) Via Pop III.2: Both the high background radiation field and the high virial temperature mean that the electron abundance is elevated above the standard value (3e-4), and so H2 formation can become extremely efficient. Obviously this requires the gas to become shielded from the Lyman-Werner radiation at some point, and it is not clear whether this actually occurs.

- Via Lyman-alpha cooling in atomic gas: LA cooling is extremely efficient (if one assumes that it is optically thin! However see Schleicher and Glover 2011), allowing essentially isothermal collapse. Very high accretion rates. Proposed as a channel for forming BH directly. Jeans mass at onset of first self-gravitating baryonic core is now very large!

- Jeans mass is very large! At a $n \sim 1$ cm^{-3} and $T = 10^4$K, $m_{\rm J} \sim 5 \times 10^6$ M$_\odot$. Supermassive BH seed?

- However unlikely that dense core will be optically thin to Lyman-Werner which would be required to keep the gas in H form. If rotation, then disc could form and perhaps H2 could form that would cool the gas down Would this lead to rapid fragmentation? Cluster, or BH?

- Even in simulations without feedback, the temperature structure of the atomic cooling halos is extremely complicated. At the outskirts there is an accretion shock where the gas is heated to the virial temperature (around 10,000 K). The rapid catalysiation of H2 due to the increased electron fraction then allows the gas to cool, creating cold streams that flow to the centre of the proto-galaxy. The chaotic nature of the temperature structure means that galaxies are born turbulent, with a wide range of Mach numbers.

- How will a BH effect such an environment. Greif et al. 2008 looked at the growth of a BH in the centre of an atomic cooling halo. They had no feedback in the model, and so they tried to gauge what the effect of a BH would be on its natal halo. First, there is the question of how quickly the BH can grow. The maximum rate at which the BH can accrete is controlled by the Eddington luminosity, giving the following "Eddington limit to the accretion rate:

$$\dot{M}_{\rm Edd} = \frac{1}{\epsilon} \frac{M_{\rm BH}}{t_{\rm Salp}}, \tag{267}$$

  where $t_{\rm Salp}$ is the Salpeter time, given by

$$t_{\rm Salp} = \frac{c\,\sigma_{\rm T}}{4\pi\,G\,m_{\rm H}} \simeq 450{\rm Myr}. \tag{268}$$

  and ϵ is the radiative efficiency (i.e. the fraction of the accretion heating that is radiated). The mass of the BH as a function of time is thus given by,

$$M_{\rm BH}(t) = M_{\rm BH,0}\,\exp\left(\frac{1-\epsilon}{\epsilon}\frac{t-t_0}{t_{\rm Salp}}\right) \tag{269}$$

  A BH can therefore enjoy exponential growth of its mass (with e-fold time $t_{\rm Salp}$), provided a suitable reservoir exists.

- We have observed 10^9 $M_\odot$ BHs at $z \sim 6$. Can we form them simply from a stellar BH seed? Naïvely, one would expect that there is no problem: a 100 $M_\odot$ BH can grow to in excess of 10^9 $M_\odot$ in less than 10^9 years – the age of the Universe at $z = 6$.

- In practise there are several problems with this. First, assuming the BHs form in 10^5 $M_\odot$ minihalos, then we've seen that ionisation feedback my be effective at shutting-off accretion onto the central objects. In addition, the BHs also may have had lower (or higher) mass companions that ended their lives in SN events. In short, stellar mass BH seeds are born starving, and have to wait until the find themselves in atomic minihalos before they can start to accrete again ($> 100Myr$).

- A further complication arises when minihalos merge (like in the formation of atomic cooling halos) – their BHs tend to merge too. This is a two-step process with the BHs first undergoing dynamical friction with the other stars, before forming a binary at the centre of the new (more massive) minihalo. As the orbit of the binary decreases due to collisions with other stellar systems, eventually the system stars to emit gravitational waves that will rapidly drain the orbit of its energy, resulting in a merger. At first this would seem to be a good thing for growing the BHs quickly, but GR predicts that the merger event is accompanied by a kick, that ejects the newly merged BH from the merger-site at a velocity > 100 km s^{-1}. As this is more than the escape velocity for young galaxies, the process of BH formation and growth must start over.

- Thus the direct collapse to a supermassive BH in an atomic-cooling halo seems the most likely way to grow the BHs to those that we see around $z = 6$.

- Greif et al. (2008) compared the accretion onto a sink particle at the centre of the atomic cooling halo to the expression for $M_{\rm BH}(t)$ above, finding that accretion rate is "super Eddington, suggesting that to follow BH formation properly, requires taking account of the feedback processes. However it also suggests that BHs could, given favourable conditions in first galaxies, accrete close to the Eddington limit for a considerable time. The BH in Getal2008 reaches a mass of 1e6 in roughly 300 Myr (and by z = 11).

- Draw Fig 13 from Greif08. Assumes $\epsilon = 0.1$.

- Even in the case of no feedback, Greif08 find that the accretion rate eventually saturates at around a few 10^{-3} $M_\odot\,{\rm yr}^{-1}$ due to the increasing amount of cold gas that flows to the centre.

- In reality, the BH is unlikely to accrete at much above the Eddington-limit, and so we can use the simple Eddington-limited accretion model to get an upper limit to the potential feedback from a BH mini-quasar. Again, we are interested in 2 features of the BH spectrum: the LW flux, and the ionising flux.

- The temperature profile of the accretion disc can be expressed as (Pringle 1981):

$$T(r) = \left(\frac{3}{8\pi} \frac{G M_{\rm BH} \dot{M}_{\rm BH}}{\sigma_{\rm SB} r^3} \right)^{1/4} \tag{270}$$

For BHs with a high spin parameter ($a \geq 0.9$), which is typical for the high-angular momentum gas in the atomic cooling halos, the inner disc (or the radius of the "inner stable circular orbit" – ISCO) is given by,

$$r_{\rm inner} \sim 2{\rm km} \left(\frac{M_{\rm BH}}{{\rm M}_\odot} \right) \tag{271}$$

One can integrate from $r_{\rm inner}$ to $10^4 r_{\rm inner}$ (after which the LW and ionisation contributions are negligible), assuming the disc radiates like a black-body, and derive the total flux in the bands of interest.

- Draw the results from Greif08 for such and analysis, showing the Q_* and $J_{\rm LW}$. They show 100 and 500 $M_\odot$ BHs, roughly the range expected from the direct collapse of massive Pop III stars (Heger & Woosley 2002). Note that the ionising photon number count in Hosokawa et al. 2008 reached few 10^{49}, while the BH count reach much higher, but only at late times – at the beginning they are infact quite similar (or even lower). The critical value of $J_{\rm LW}$ for the suppression of H_2 formation in halos is around 10^{-2}, which is easily exceeded by the BH. This suggests that BHs may be able to prevent H_2 and HD cooling from becoming important in atomic cooling halos.

- In conclusion, the formation of BHs due to high accretion rates makes it likely that a halo will be forced to form stars via the Lyman alpha cooling model – favouring the formation of even more BHs.

9 The onset of Pop II star formation

9.1 The definition of Pop II star formation?

- The definition of Pop III and Pop II would seem to be clear – either we have simply a 'primodial' distribution of metals in the case of Pop III (which essentially amounts to a trace abundance of Li), or we have an elevated amount of metals, in which case the star is Pop II. However if the amount of metals in the gas is too low to alter the way in which the stars are formed, then can those stars still be thought to be primordial, or metal enriched?

- A more practical definition is too ask at what point does the metallicity begin to affect the way in which stars form. Or more precisely, at what metallicity does the IMF shift from that dominated by high-mass stars in the Pop III case, to that dominated by low-mass stars in the Pop II/I case.

- Which brings the question: what is the IMF of primordial stars? Currently unknown. From what we've seen so far, the mass scale seems to be larger than in present-day stars, so we assume that the characteristic mass (either the mean or the median) is larger than we see today. However the shape of the IMF is unknown: often assumed to be peaked at some high mass, and then either log-normal in shape, or with a Salpeter-like

power-law tail at the high-mass end. From the simulations (which have only covered an extremely short period of the star's life), it seems that the IMF may be flat. Draw these different IMFs on the board.

- The current research focuses on a fairly simple question: at what metallicity does the gas start to fragment at mass-scales smaller than in Pop III. More importantly, exactly what feature of the enriched gas is responsible for the cooling. Before looking at this, we will first take a brief look at how the gas is enriched via SN explosions.

9.2 Supernovae enrichment

- Atomic cooling halos have a potential energy of $\sim 10^{53}$ erg, and so the baryons can easily survive SN (even though their constituent sub-halos may be emptied). This is another reason why they are considered 'galaxies' - they are the first objects that can chemically self-enrich.

- Given the difference in the formation time between the formation of the first Pop III forming halos and the atomic cooling halos, there is plenty time for SN, even from the lower mass (15-40 Msun) progenitors.

- To date, only the higher-mass progenitors have been studied (i.e. those with masses 140 to 260 Msun – the PISN). Lower mass progenitor studies are currently underway.

- PISN have also very high yields. The central core of PISN can range from 65 to 130 $M_\odot$, and is enriched to at least solar composition in C and O, if not much above. As we will discuss, these are the important coolants in metal enriched gas, so we will focus our discussion on how they are distributed. Rather than using the old $[Fe/H]$ notation to describe the metallicity, it now more common to describe the abundance of each element with respects to the equivalent solar value. A gas with metallicity $Z_\odot$ would have all metals in equal abundance to those in solar composition gas. A gas with 0.1 $Z_\odot$ would have only a tenth as much C, Si, O etc as standard solar type gas.

- Greif10 looked at the distribution of metals from a Pop III PISN with progenitor mass ~ 200 (at the point of SN explosion) and central engine of 10^{52} erg. Since a star of this mass has 100 $M_\odot$ of metals (essentially C, Si, O), they assign 0.5 times the solar abundance to the 200 $M_\odot$ at the centre of each collapsing minihalo, and inject the energy into kinetic motions. The area around the SN has already been cleared by the ionising and dissociating flux from the parent star, and so the SN is free to expand into the inter-halo medium (IHM), as the first galaxy is assembled.

- Their results suggest that after 300 Myr (around z = 13), the IHM has a metallicity of around $10^{-4}\, Z_\odot$. This implies that the first galaxies formed with pre-enriched gas.

- However, from these low resolution calculations, it seems that the metallicity of the high density peaks remains low - if not zero: they start to collapse before the SN/feedback becomes important. Draw Fig 7. from Greif2010. So although the effect of SN and

ionisation feedback is dramatic, the sites of star formation may be largely unaffected. See also Whalen et al. (2008a,b; 2009).

- Note that this may not hold as the resolution increases. Recent results by Whalen have demonstrated that relic HII regions can produce significant amounts of molecular hydrogen, due to the large free electron abundance. This can create a cool layer inside halos, that becomes buoyant, and could drive further mixing.

9.3 Cooling by metals

- The goal is to define a critial metallicity, Z_{crit}, above which the gas can cool more effectively than it can in the pure primordial case.

- The most effective coolants are the CII and OI (that is, singly ionised C, and neutral O). They are also the most abundant species after He, and have the highest yields from SN. They emit via 'forbidden' transition lines (that is, from long-lived states). As these are the most effective and abundant species, much of the research has focused on determining at what metallicity these can be come important coolants.

- As we discussed in previous lectures, the basic condition for fragmentation is $\Lambda_X \geq \Gamma_{pdV}$, where Λ_X is the cooling (erg s^{-1} cm^{-3}) for species X. Bromm & Loeb (2003) looked at the cooling provided by CII and OI, and derived critical metallicities for the two species by finding when,

$$\Lambda_{CII,OI}(n,T) \simeq \frac{3}{2}\frac{n\,k_B\,T}{t_{ff}}. \tag{272}$$

- Clearly we have a dependence on the temperature and density in this relation, so what values do we use? BL03 opted for T $\sim$ 200 K and $n \sim 10^4 cm^{-3}$, since these are the conditions at the point when H_2 starts to become inefficient in the standard Pop III(.1) case. As such, if cooling provided by either CII or OI is greater than the pdV heating at this point in the collapse, it would give rise to a reduction in the characteristic mass for the fragments.

- Considering C and O cooling separately, BL03 found critical metallicities for the two species of $Z_{C,crit} = 10^{-3.5} Z_{C,\odot}$ and $Z_{C,crit} = 10^{-3.05} Z_{O,\odot}$. This implies that around a metallicity of $Z = 10^{-3}\,Z_\odot$, the is a transition to a new regime of star formation, that is driven by metal-line cooling, and has smaller characteristics masses than found in standard Pop III.

- Note that in the case of CII, there is obviously the assumption that all (or at least the majority) of the carbon has been ionised. BL03 reckon this is a good approximation as the soft UV of 11.26 eV can keep the C ionised. And since the C abundance is low (*solar* has $x(C) = 1.4 \times 10^{-4}$), these photons should have no problem penetrating the ISM.

- Clearly the two coolants can also act together, so Frebel et al. (2007) introduced a new concept of a "transition discriminant" which defines the total amount of C + O needed to "transition" to a new IMF. Again, we're trying to satisfy the condition,

$$\Lambda_{\rm CII} + \Lambda_{\rm O} \geq \Gamma_{pdV}. \tag{273}$$

At T $\sim$ 200 K and $n \sim 10^4 {\rm cm}^{-3}$, the cooling can be roughly written as (Staller & Palla 2005),

$$\Lambda_{\rm CII} \simeq 6 \times 10^{-20} {\rm ergs}^{-1} {\rm cm}^{-3} \left(\frac{n_{\rm C}}{n_{\rm H}}\right) / \left(\frac{n_{\rm C}}{n_{\rm H}}\right)_\odot \tag{274}$$

$$\Lambda_{\rm OI} \simeq 3 \times 10^{-20} {\rm ergs}^{-1} {\rm cm}^{-3} \left(\frac{n_{\rm O}}{n_{\rm H}}\right) / \left(\frac{n_{\rm O}}{n_{\rm H}}\right)_\odot \tag{275}$$

and the heating is given simply by,

$$\Gamma_{pdV} \simeq 2 \times 10^{-23} {\rm ergs}^{-1} {\rm cm}^{-3} \tag{276}$$

The condition above can then be written as,

$$\left(\frac{n_{\rm C}}{n_{\rm H}}\right) / \left(\frac{n_{\rm C}}{n_{\rm H}}\right)_\odot + 0.3 \left(\frac{n_{\rm O}}{n_{\rm H}}\right) / \left(\frac{n_{\rm O}}{n_{\rm H}}\right)_\odot \geq 0.3 \times 10^{-3} \tag{277}$$

or

$$10^{\rm [C/H]} + 0.3 \times 10^{\rm [O/H]} \geq 10^{-3.5} \tag{278}$$

since,

$$\rm [X/H] = \log_{10} \left[\frac{n_X}{n_H}\right] - \log_{10} \left[\frac{n_X}{n_H}\right]_\odot \tag{279}$$

This allows Frebel et al 2007 to define the transition discriminant as,

$$D_{\rm trans} \equiv \log_{10}(10^{\rm [C/H]} + 0.3 \times 10^{\rm [O/H]}) \geq -3.5 \pm 0.2 \tag{280}$$

where the ± 0.2 arises from variations in n and T.

- Draw Fig 1. from Frebel et al. 2007.

- Finally, note that the CMB temperature provides a limit to how effective the metal cooling can be: even if there is a quick transition to high metallicities at high redshifts, the CMB temperature will regulate the mass of the objects that can form in this regime. The fragment masses then may be not much lower than in the HD cooling case (i.e. Pop III.2).

9.4 Cooling by dust

- An alternative route by which the gas can cool is via collisions with dust grains. The idea is fairly simple: the gas transfers its kinetic energy to the grain during inelastic collisions, heating up the dust grain in the process. The dust then radiates away

the energy, behaving approximately as a blackbody, with the result that much of the radiation is lost from the cloud. This particularly effective cooling process is responsible for maintaining the low temperatures of pre- and proto-stellar cores in present-day star formation, and is able to dissipate the $p\mathrm{d}V$ work associated with their collapse to form stars.

- The recent paper by Schneider et al. (2012) gives a good account of this processes, and we follow their arguments (and notation) here.

- The ability of the dust to act as a coolant is controlled by two processes. The first process is the dust's ability to radiate away its thermal energy. The thermal dust emission rate can be expressed as,

$$\Lambda_{\rm gr} = 4\sigma_{\rm SB} T_{\rm gr}^4 \kappa_{\rm P} \beta_{\rm escp} \rho_{\rm gr}. \tag{281}$$

  Note that density of grains can be expressed in terms of a dust-to-gas ratio D, via $\rho_{\rm gr} = D\rho$. The Planck mean opacity is given by,

$$\kappa_{\rm P} = \frac{\pi}{\sigma_{\rm SB} T_{\rm gr}^4} \int_0^\infty B_\nu(T_{\rm gr}) \kappa_\nu \mathrm{d}\nu \tag{282}$$

  and the photon escape probability $\beta_{\rm escp}$ is given by,

$$\beta_{\rm escp} = \min(1, \frac{1}{\tau^2}) \tag{283}$$

  where the optical depth is given by $\kappa_\nu \rho \lambda_{\rm J}$ and λ_J is the Jeans mass. The escape probability takes into account that at some density, the dust is going to become optically thick to its own radiation and will be unable to freely radiate away the compressional heating.

- The second process is the transfer of the kinetic energy of the gas to the dust grains. Note that the energy transfer can go both ways, i.e. from the gas to the dust and vice-versa, depending on the relative temperature of the gas and dust. Hollenbach & McKee showed that the heating of the dust by the gas can be expressed via,

$$H_{\rm gr} = \frac{n_{gr}\,(2k_{\rm B}T - 2k_{\rm B}T_{,\rm gr})}{t_{\rm coll}} \tag{284}$$

  where T and $T_{,\rm gr}$ are the gas and dust temperature respectively, and $t_{\rm coll}$ is the timescale for collisions between gas components and the dust grains, given by $(n_{\rm H}\sigma_{\rm gr}\bar{v}_{\rm H} f)^{-1}$. Here, $n_{\rm gr}$ is the number density of grains, $\sigma_{\rm gr}$ is their cross-section, and $\bar{v}_{\rm H}$ is the mean velocity of the hydrogen (the main collider). We assume the dust grains are significantly heavier than the hydrogen and can be taken to be stationary, and that the hydrogen follows a Maxwellian velocity distribution,

$$\bar{v}_{\rm H} = \left(\frac{8k_{\rm B}T}{\pi m_{\rm H}}\right)^{1/2}. \tag{285}$$

We can relate the dust properties to the gas properties by making use of two parameterisations of the dust. First we define S, the grain cross-section per unit mass of dust. Second we define D to be the dust-to-gas ratio (for present-day, solar metallicity gas, this is around 0.01). This permits us to write the product $n_{\rm gr}\sigma_{\rm gr}$ that appears in the numerator as $n_{\rm H}m_{\rm H}\mu\, SD$.

- The dust temperature can be found by equating the heating rate of the dust by collisions with the gas, and the rate at which the dust can radiate, $H_{\rm gr} = \Lambda_{\rm gr}$ and solving for the temperature. In equating these two relations we find that the grain temperature becomes independent of D, depending only on T, ρ, S and the values for the opacity, κ_P.

- As the density increases during the collapse, the gas transfers more and more of its energy to the dust, increasing the dust temperate but also increasing the rate at which the dust can radiate away the energy. Draw this on the board for different metallicities.

- Schneider et al. derive a minimum amount of dust that is required for fragmentation, in terms of the dust to gas ratio D. To do take the requirement that the rate of grain heating is equal to the compressional heating term $H_{\rm gr} = \Gamma_{\rm pdV}$, such that all of the energy created by collapse is transferred to the grain, where is assumed to be radiated away. If one then assumes that $T_{\rm gr} << T$, one can derive,

$$SD > 1.4 \times 10^{-3}{\rm cm}^2{\rm g}^{-1} \left(\frac{T}{10^3{\rm K}}\right)^{-1/2} \left(\frac{n_{\rm H}}{10^{12}{\rm cm}^{-3}}\right)^{-1/2} \tag{286}$$

  Assuming S to be roughly $3.5 \times 10^5{\rm cm}^2\,{\rm g}^{-1}$, then we find

$$D_{\rm crit} > 4 \times 10^{-9}{\rm cm}^2{\rm g}^{-1} \left(\frac{T}{10^3{\rm K}}\right)^{-1/2} \tag{287}$$

  The dust to gas ratio and the metallicity are related through the depletion factor $f_{\rm dep} = M_{\rm dust}/(M_{\rm dust} + M_{\rm met})$, by $D = f_{\rm dep}\, Z$. Note here that Z is in absolute units, where $Z_\odot \sim 0.02$. Although SN are predicted to convert most of their metals into the form of dust, much of this can be destroyed in the 'reverse shock' that forms as the interior cavity starts to cool. The parameter $f_{\rm dep}$ is therefore useful, as it can be used to describe the effects of the reverse shock on the metal content. For solar-type depletions factors of around 0.5, the above value of $D_{\rm crit}$ would predict fragmentation whenever the gas metallicity is $> 10^{-6}\, Z_\odot$.

- Indeed, in their one-zone models, Schneider et al do find that the the gas cools faster than it heats whenever D is above the critical value, but not when the D is lower. They also express this in terms of the range of depletion factors for a given metallicity.

- We find somewhat higher metallicities in the fully 3D turbulent simulations, of around 10^{-5} to $10^{-4}\, Z_\odot$ (assuming scaled-down solar composition dust and gas). This is due to the spread in the rho-T diagram due to turbulent motions and shocks that effectively washes out the small dip in the temperatures provided by the dust at lower metallicities.

- Regardless, the dust cooling suggests a transition to more fragmentation, and hence a different IMF at much lower metallicities than we seen in the metal-line cooling case.
- However, note that this fragmentation occurs at extremely high densities compared to those we discussed in the case of the HD and metal-line cooling. More similar to the fragmentation found in Pop III disc case.
- Finally, note that the CMB temperature plays much less of a role here, since the temperatures at the bottom of the dip are typically much higher than they are in the metal-line cooling case. The main parameter dominating the Jeans mass is the density, in the case of dust cooling. The CMB only becomes important at very high metallicities, when the dip can be large.

9.5 The IMF: transition from Pop III to Pop II

- Focus on the Omukai plot and show the two cooling regimes.
- We see that the rho-T plot now has two clear regimes of fragmentation: one induced by the onset of metal-line cooling and the other by dust cooling. Furthermore, they are clearly separated in density space. So which one is more important?
- Based on a pure Jeans mass argument, we see that the dust cooling is more likely to produce low-mass objects than the metal-line cooling.

9.6 Evidence from the observations

- The odd-even pattern predicted by PISN is not observed in the very metal poor stars discovered so far. Favours a Pop III IMF that was dominated by CCSN. Draw the odd-even pattern on the board.
- Summarise Frebel et al (2007). Shows that most of the stars are above the transition discriminant and are therefore supports the line-cooling idea. Those stars which are under the transition discriminant, in terms of [Fe/H], are often found to high very high [C/H] or [O/H], and so are not as metal poor as their [Fe/H] value suggests.
- Most metal poor stars have low masses – considerably below the Jeans mass at the dip in the Omukai plot for the metal-cooling regime. If these stars were representative of the IMF at that metallicity, then they imply that metal cooling alone could not have been responsible for setting their mass.
- However, the most conclusive evidence however comes from the 'Leo' star, discovered by Caffau et al. 2012 (the group at the LSW), which has a metallicity of $Z = 10^{-5\pm1}$. As this is considerably below the transition discriminant, in the "forbidden zone". It implies that dust-cooling was a more likely scenario for its formation. If that is true, it implies that the $f_{\rm dep}$ was at least 0.01 at $Z = 10^{-5}\,Z_{\odot}$.

10 Reionisation

10.1 Evidence for reionisation

- Perhaps the best evidence for the existence of a mainly ionised inter-galactic medium (IGM) is via the spectra of bright, high-redshift quasars. Quasar emission is associated with the extremely hot environment that surrounds a SMBH. The regions around the quasar are able to emit strongly over a wide range of frequencies, including those that coincide with the Lyman-series of hydrogen.

- If neutral hydrogen is present along the line of sight between us and the quasar, then it can absorb the Lyman-series emission from the quasar. Now imagine the quasar is very distant, at some redshift z, such that its spectra is shifted to wavelengths $\lambda = \lambda_0(1+z)$. Now, local clouds of neutral H in the Milky Way can absorb radiation that originated at much shorter wavelengths, but that have been red-shifted into one of the Lyman-series bands by cosmic expansion. This process is obviously not just limited to the MW: all HI clouds along the line of sight to the quasar can absorb any photon that arrives at the Lyman-bands. The effect is particular strong in the Lyman-α band (1216 Å), and gives rise to the 'Lyman-α forest'. Draw this on the board.

- So how much gas is required to completely absorb the Lyman-αemission from distant sources? To figure this out, we need to look at the effective optical depth for Lyman-α, given by

$$\tau_{\mathrm{L}\alpha} = \frac{\pi e^2 f_\alpha \lambda_\alpha n_{\mathrm{HI}}(z)}{m_{\mathrm{e}} c H(z)} \tag{288}$$

  where $H \approx 100h\mathrm{km\,s^{-1}Mpc^{-1}}\Omega_m^{1/2}(1+z)^{3/2}$, and so

$$\tau_{\mathrm{L}\alpha} \approx 6.45 \times 10^5\, x_{\mathrm{HI}} \left(\frac{\Omega_b h}{0.0315}\right) \left(\frac{\Omega_m}{0.3}\right)^{-1/2} \left(\frac{1+z}{10}\right)^{3/2} \tag{289}$$

  for a matter dominated Universe. We see that the optical depth to the Lyman-α line is extremely high, requiring only a small amount of HI (as expressed via the abundance x_{HI}) to reach unity. At a redshit of around $z \sim 6$, we get $\tau_{\mathrm{L}\alpha} \sim 1$ for $x_{\mathrm{HI}} \sim 10^{-5}$. Clearly this is much smaller than we found in the Pop III star forming minihalos, and so it expected that much of the Universe will be able to completely absorb distant quasar light. The fact that we do see *some* emission red-wards of Lyman-α peak in the quasar spectrum tells us that the Universe is extremely well ionised. The 'forest' simply represents the dense HI clumps than sit in an otherwise completely ionised Universe. This is perhaps the most unambiguous proof that we live in an ionised Universe.

- However, if the quasar was distant enough, then we should see the complete absorption of the Lyman-α line above the redshift at which the Universe becomes ionised. This is referred to as the Gunn-Peterson trough. Draw on the board. In 2003, Fan et al. found the first quasars to show the Gunn-Peterson trough, and they had redshifts of around 6 or greater.

- Note however that there is some scatter: not all the $z > 6$ quasars present a GP-trough, while some quasars just below $z = 6$ do. Suggests that the reionisation of the Universe was not homogeneous, but rather patchy. As we go to higher redshifts, the number of quasars found obviously drops, and it becomes difficult to say much more.

- So we good evidence that the Universe was, on average, fully ionised by around $z = 6$, but when did it start?

- A further constraint on the onset of reionisation comes from the CMB. Clearly if the gas is ionised at high redshift, then electrons can scatter the CMB photons that have been unimpeded since recombination, leaving a signal in the present-day CMB. To calculate the effect, it is common define a visibility function,

$$g\eta = -\dot{\tau} e^{-\tau(\eta)}, \tag{290}$$

where η is the conformal time ($\equiv \int \mathrm{d}t/a$), and $\dot{\tau} = \mathrm{d}\tau/\mathrm{d}\eta$. This visibility function gives the probability that a CMB photon has been scattered out of the line of sight between η and $\eta + \mathrm{d}\eta$. The optical depth to for Thompson scattering is

$$\tau(\eta) = -\int_{\eta}^{\eta_0} \mathrm{d}\eta \dot{\tau} = -\int_{\eta}^{\eta_0} \mathrm{d}\eta\, a(\eta) n_{\mathrm{e}} \sigma_{\mathrm{T}}, \tag{291}$$

where η_0 is the present time. One can then integrate along each line of sight to estimate the suppression in the temperature fluctuation due to the epoch of reionisation (i.e. how long, in terms of z it has persisted). Clearly since the effects are expected to be small, the optical depth is also expected to be small. Further constraints come from the polarisation of the map: any additional scattering at $z < 1068$ will cause polarisation on scales larger than the recombination horizon scale.

- From the WMAP observations, the implied that reionisation started around z_{i} is 11 ± 1.4, and was finished around $z = 7$.

10.2 How was the Universe is reionised?

- In the simplest argument, one can say that the Universe is reionised when there exists roughly 1 ionising photon per baryon. Obviously, this is a clear minimum, since it assumes that recombinations are not important, however it does yield an order of magnitude estimate that provides a useful insight into what is obviously a complicated process.

- Central to the argument is the concept of the number of ionising photons emitted per baryon, which we will denote by A. If one consider a single star, then the definition is simply,

$$A = \frac{\text{total number of } > 13.6\text{eV photons (over life)}}{\text{number of baryons in star}} \tag{292}$$

Obviously, the denominator is simply M_*/m_{p}. In principle the numerator is also fairly simple – one simply integrates the stellar spectrum over the life of the star from the

Lyman-limit upwards – however the situation is complicated by the fact that full evolution of the star is often uncertain.

- One can also define A for an IMF (see Ciardi & Ferarra 2005 for a summary), which involves integrating over stars of different masses and needs to account for the different ages of the stars, and the different stages of evolution. For non-stellar sources, such as BH and QSOs, then one needs to consider how these objects can accrete, and so A becomes a strong function of the star formation rate.

- For very massive Pop III stars ($> 140\,\mathrm{M}_\odot$), however, the numbers are fairly well constrained, with $A \sim 10^5$. For a single population of massive Pop III stars, the total number of ionising photons is then given by,

$$N_{\rm ion} = A_{\rm PopIII}\, f_{\rm SFE}\, N_{\rm Bary} \tag{293}$$

where $N_{\rm Bary}$ is the total number of baryons in the Universe, and $f_{\rm SFE}$ is the fraction of those baryons that have been converted into stars. If one then sets $N_{\rm ion} = N_{\rm Bary}$, then we find that the star formation efficiency in Pop III required to reionise the Universe is,

$$f_{\rm SFE} = \frac{1}{A_{\rm PopIII}} \sim 10^{-5} \tag{294}$$

So Pop III stars could reionise the Universe with a very small SFE!

- Alternatively we can look at the effects of Pop II star formation. As the majority of the mass in a standard IMF is locked up in low-mass stars – which have very little ionising flux – the value of A is much lower, at roughly ~ 4000. Pop II stars also form in galaxies, so it makes more sense to talk about the fraction of mass in the galaxy that has been turned into stars, which we will denote as f_*. We then also have to define the fraction of the mass in the Universe that has collapsed into galaxies, $f_{\rm coll}$. Finally, we should also consider that many photons may not reach the IGM, but instead become locked up in the high density regions within the galaxy, and do we need to introduce $f_{\rm esc}$, the fraction of the photons that escape from the galaxy.

- From present-day star formation we see that roughly 10% of the mass in a star forming region is converted to stars. The escape fraction is also estimated to be around 10% (more on this below). If we adopt these values, then the total fraction of the Universe required to be in collapsed structures for reionisation is

$$f_{\rm coll} = \frac{1}{f_{\rm esc}\, f_*\, A_{\rm PopII}} \sim 2.5\% \tag{295}$$

- However, as noted the above estimates ignored recombinations. The recombination timescale is given by $\langle t_{\rm rec}\rangle = (\alpha_{\rm B} n_{\rm e} C_{\rm HII})^{-1}$. The factor $C_{\rm HII} = \langle n_{\rm HII}^2\rangle / \langle n_{\rm HII}\rangle^2$ is mean effective clumping factor of the ionised gas. If the recombination timescale is greater than the Hubble time, $t_{\rm H}$, then the total number of photons per baryon required to reionise the Universe is given by,

$$N_{\rm ion}/N_{\rm Bary} \sim \max\{t_{\rm H}/\langle t_{\rm rec}\rangle, 1\} \tag{296}$$

where

$$\langle t_{\rm rec}\rangle \sim 1.7\,{\rm Gyr}\left(\frac{\Omega_b h^2}{0.02}\right)\left(\frac{1+z}{7}\right)^{-3} C_{\rm HII}^{-1} \tag{297}$$

$$t_{\rm H} \sim 1\,{\rm Gyr}\left(\frac{1+z}{7}\right)^{-3/2}\left(\frac{h}{0.7}\right)^{-1} \tag{298}$$

So for the Einstein-de Sitter Universe, $\langle t_{\rm rec}\rangle < t_{\rm H}$ for redshifts less than around 9, and $N_{\rm ion}/N_{\rm Bary}$ required for ionisation becomes larger than 1, and much earlier if the gas is predominantly in dense clumps.

- Obviously, the ionising photons come at a price: metals. The question is then, for a given stellar population, does that accompanying metal enrichment required for reionisation match the observations. Ricotti & Ostriker (2004) looked at this and found that even under the most favourable conditions, such as over-estimating A and underestimating the metal yield), the accompanying metal enrichment from pure Pop III stars would be extremely high, at around $0.001{\rm Z}_\odot$ in the IGM and up to ${\rm Z}_\odot$ in the over-dense regions.

- These results suggest that BH accretion may have played a stronger role in the early stages of reionisation, or that PISN were not that common. Note that as we go to lower masses of Pop III stars, value of A does not drop by that much, while the total metal yield does. Either way, the results suggest that Pop II star formation will be well underway by the end of reionisation.

- So how does the ionisation process proceed? The ionisation starts on the smallest scales, as we have already discussed above: individual Pop III stars ionise the region around them, with gradually larger and larger scales becoming ionised. Initially the Universe is characterised by isolated regions of ionisation, until these eventually merge.

- Draw the classic sketch of the Universe as a function of redshift.

- An interesting feature of the ionisation fronts is that they behave different from those in the classic interstellar medium once they have broken out of their collapsing overdensities and reached the scales on which the Hubble flow dominates. The comoving Strögren radius is given by

$$r_{\rm S}(t) = \left[\frac{3Q_*}{4\pi\alpha_{\rm B} n_H^2}\right]^{1/3} a(t) \equiv R_{\rm S}\, a(t), \tag{299}$$

in keeping with the notation that we used earlier in the course. If the ionisation was to occur very rapidly, then $R_{\rm S}$ is steady state solution, and then $r_{\rm S}(t)$ would describe how this region evolves as the Universe expands. A detailed analysis, as given by Shapiro & Giroux ('87) shows that picture is actually more complicated than this, and the ionisation region never achieves the balance that is assumed in the Strömgren volume approach, as $r_{\rm S}$ evolves faster than the ionisation front.The front therefore remains R-type, never driving a pressure induced shock into the IGM.

- This means that the ionisation regions created by the sources remain frozen in the volume of space, never quite reaching their maximum extent. So for the pocket of ionisation to merge with one another requires either that new ionisation regions appears (until all the volume is filled), or for the central engine to grow. In practice both these processes occur.

- The large-scale topology of reionisation is 'inside-out', in the sense that the dense regions ionise first, while the underdense voids only reionise at the end of the reionisation process. Although the denser regions are more difficult to ionise, since recombinations are more efficient there, they contain many more ionising sources than their lower density counterparts: a region with a density enhancement only 10 % above the mean of the Universe can have a 50 % higher concentration of galaxies.

- Once the ionisation bubbles meet, the ionisation background increases sharply as there is now an excess of ionising photons for much of the Universe's volume. This changes the Jeans mass in the baryons, and severely limits the mass of the halos in which star formation can proceed. Before reionisation, the IGM is cold (tens of K) and neutral, and so the Jeans mass plays a secondary role to cooling in the formation of bound star-forming regions: the dark matter will pull the gas in, and it will eventually start to cool once H_2 formation kicks in, which drops the Jeans mass. After reionisation, the Jeans mass is increased by several orders of magnitude to the point where halos which previously could form stars, are now unable to drag the gas in.

- Including the gravitational effect from the DM, the ionisation results in a linear Jeans mass corresponding to a halo circular velocity of

$$V_J \approx 80 \left(\frac{T_{\rm IGM}}{1.5 \times 10^4 {\rm K}} \right)^{1/2} {\rm kms}^{-1} \tag{300}$$

  Recall that the circular velocity is given by,

$$V_c = \left(\frac{GM}{r_{\rm vir}} \right)^{1/2} = 23.4 \left(\frac{M}{10^8 h^{-1} {\rm M}_\odot} \right)^{1/3} \left[\frac{\Omega_m \delta_c}{18\pi \Omega_m^z} \right] \left(\frac{1+z}{10} \right)^{1/2} {\rm kms}^{-1} \tag{301}$$

  Clearly we see that the ionisation of the IGM is able to inhibit the formation of lower mass galaxies. In halos with $V_c > V_J$ the fraction of infalling gas equals the universal mean (Ω_b/Ω_m), but in halos below this velocity, the accretion rate is strongly suppressed.

- Of course, once the IGM is ionised, it needs to be maintained. Assuming that the bulk of the work is done by Pop II stars at $z < z_{\rm re-i}$, then the star formation rate per unit comoving volume required to balance the recombinations is given by,

$$\dot{\rho}_* \approx 2 \times 10^{-3} f_{\rm esc}^{-1} C \left(\frac{1+z}{10} \right)^3 {\rm M}_\odot {\rm yr}^{-1} {\rm Mpc}^{-3}. \tag{302}$$

  From the observations, it would seem that this condition is met, and so the ionisation of the Universe can be maintained by the current star formation.